USES AND ABUSES OF INTELLIGENCE

Studies Advancing Spearman and Raven's Quest for Non-Arbitrary Metrics

JOHN AND JEAN RAVEN
(editors)

Royal Fireworks Press, New York
Competency Motivation Project, Edinburgh
EDGE 2000 Ltd., Budapest
RTS Romanian Psychological Testing Services SRL, Romania

PART I: Introduction to the *Raven Progressive Matrices* Tests: Conceptual Basis, Measurement Model, and a Few Findings

PART II: Practical Measurement Issues: Lessons from 75 Years' Work with *Item Response Theory*: Benefits, Problems, and Potential Solutions

PART III: Stability and Change in RPM Norms Across Time and Culture

PART IV Outstanding Conceptual and Measurement Issues

PART V: Emerging Applications

PART VI: Some Outstanding Ethical Issues

Individual chapters are also available as separate PDF files at
http://eyeonsociety.co.uk/resources/fulllist.html#uses_and_abuses

PART I

Introduction to the Raven Progressive Matrices Tests: Conceptual Basis, Measurement Model, and a Few Findings

The first chapter in this Part of our book initially sets out to clarify what it is that Raven's *Progressive Matrices* (RPM) and *Vocabulary* (MHV) tests seek to assess and the theoretical and psychometric principles deployed in their development. Thereafter, it briefly summaries research (discussed more fully in later chapters) dealing with such things as heritability, changes in scores over time, and the stability in norms across cultural groups. These data prompt a re-examination of research which has contributed to a number of serious myths and misunderstandings.

In the second chapter, by demonstrating that a series of Piagetian tasks behave as if they were widely spaced *Progressive Matrices* items, Irene Styles provides remarkable evidence confirming the "existence" of "general cognitive ability" and its measurability via a series of items of increasing complexity and conforming to the requirements of Item Response Theory.

Chapter 1

General Introduction and Overview: The *Raven Progressive Matrices* Tests: Their Theoretical Basis and Measurement Model

John Raven*

Introduction

Some readers may be so familiar with Raven's *Progressive Matrices* tests that they will be tempted to regard this chapter as redundant.

It has, however, been our experience that many users of the tests have not fully understood what they were designed to measure, still less the measurement model used to develop them. This has led to widespread misapplication of the tests in both research and practice and to extensive misinterpretation of research results.

Most of the chapters of this book present the results of relatively recently completed research, both substantive and methodological. Much of this material will be new to many readers. Nevertheless its true significance will be lost on those who have in the past sought to impose what might be called a classical theoretical (interpretational) framework and measurement model on the *Progressive Matrices.*

For these reasons, we would encourage most readers to at least skim through this chapter, allowing themselves to read more deeply when some topic catches their eye.

In addition to outlining what the *Raven Progressive Matrices* (RPM) tests set out to do and the measurement model behind them, the chapter briefly summaries research dealing with changes in scores over time (and the way in which that research prompted the development of a new

* The author is indebted to very many people for material included in this chapter but especially to his wife, Jean Raven, for making endless alterations to the Figures and Tables, to Joerg Prieler for the IRT based analyses, and to Michael Raven for generating the illustrative items shown in Figures 1.1 to 1.6.

version of the *Standard Progressive Matrices* – the SPM ***Plus***) and the stability in the norms across cultural groups. Although these data will be discussed more fully in later chapters, they have made it necessary to re-evaluate a great deal of research – often conducted on inadequate samples – which has contributed to serious myths and misunderstandings over the years. It is therefore important to try to correct some of these misunderstandings as quickly as possible.

The Raven Progressive Matrices Tests and Their Philosophy

It is perhaps easiest to introduce Raven's *Progressive Matrices* tests by discussing a couple of items similar to those of which the tests themselves are composed. When taking the tests, respondents are asked to select the piece needed to complete patterns such as that shown in the upper part of Figures 1.1 and 1.2 from the options in the lower parts of the Figures.

Figure 1.1 **An "Easy" *Standard Progressive Matrices* Item**
(similar to one in the Test itself)

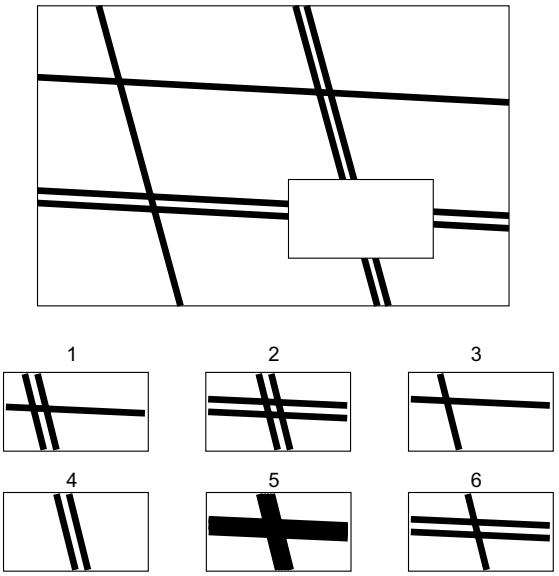
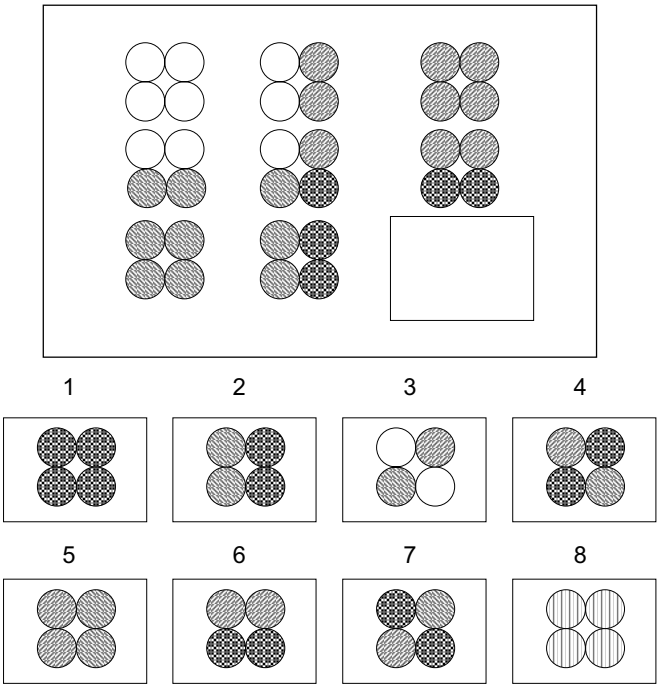
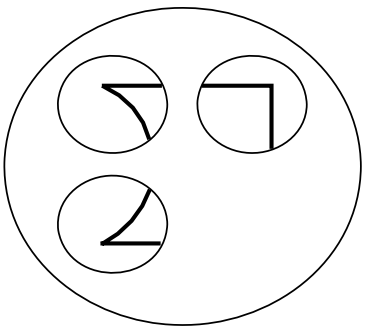


Figure 1.2 **A Moderately Difficult *Standard Progressive Matrices* Item**
(similar to one in the Test itself)



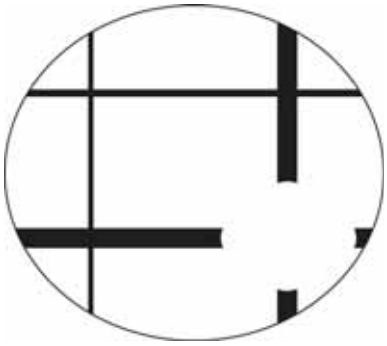
But, to illustrate what the tests are really trying to measure, it is useful to consider the "simpler" items shown in the next four Figures.

Figure 1.3



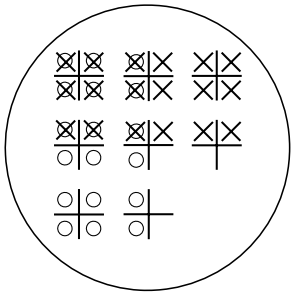
WHAT?

Figure 1.4



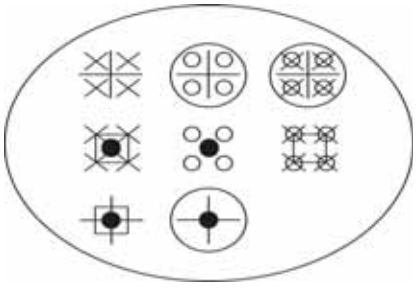
WHAT?

Figure 1.5



WHAT?

Figure 1.6

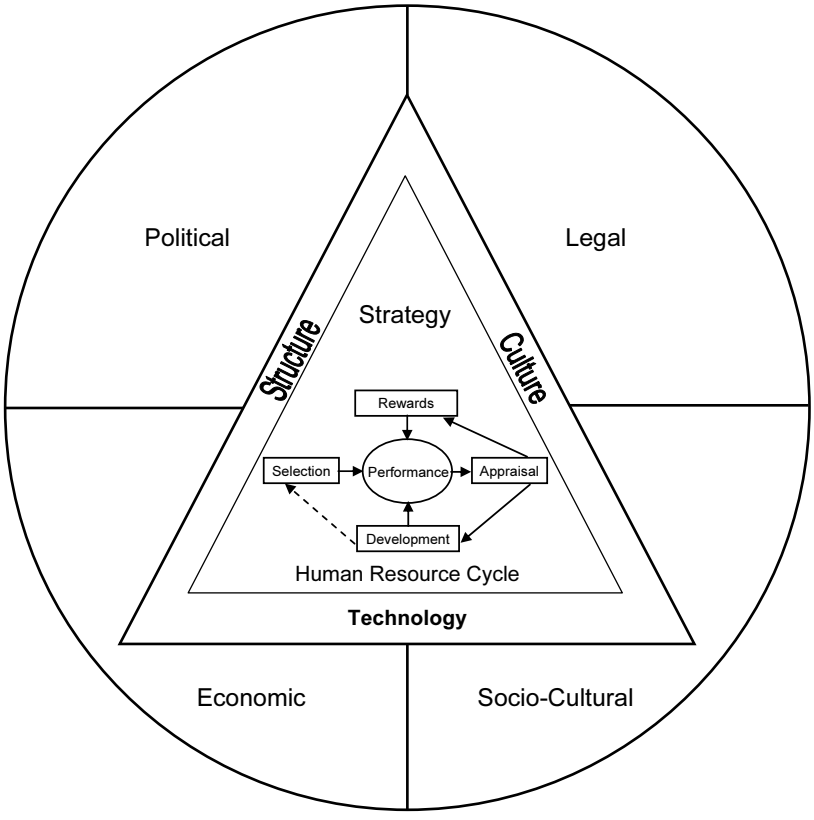


WHAT?

Note that Figures 1.3 to 1.6 have not been presented as "problems". They may be regarded as problems. Or they may not be. If you do regard them as presenting a problem your question might be "What is it?" or "How does it work: what's the logic?"

Now let us consider a related picture. Suppose you are a manager and you are interested in the success of your business. You are wondering how to move forward. You are thinking about the context of your business. What to make of the external social, economic, political, and cultural context which so much determines its success and how to intervene in it. In fact, you are pondering the situation depicted below.

Figure 1.7



WHAT?

Reproduced, with permission, from Lees (1996)

There is very little to guide you in thinking about the field of forces depicted in Figure 1.7 – how to think about such things as how to develop and utilise subordinates' motives and talents (human resources), how to think about the external economic and social processes that so much determine the success of your business, how to harness those external forces (as one uses the sails of a boat to harness the wind), where they are likely to push you if you do not understand them, what new opportunities they open up … and so on.

So, what do you *see* as a manager?

It is the strength of people's desire to make sense of such "booming, buzzing, confusion" (and their ability to do so) that Raven's *Progressive Matrices* (RPM) tests set out to measure – and, as we shall see, to some extent, do measure.

Some people have difficulty seeing even the rectangle in Figure 1.4. Some see immediately the whole design in Figure 1.6 and its internal logic. Most people need to look at the design in Figure 1.6 and ask "What is it?"; "What might it be?"; "Does this part here tell me anything about what the whole might be?"; "Does this glimmering insight into what the whole might be tell me anything about the significance of this *part*?"

More specifically, the tests set out to measure meaning-making – or, more technically, 'eductive' – abiltiy.

This involves the use of feelings to tell us what we might be looking at; which parts might be related to which other parts and how. Which parts beckon, attract, give us the feeling that we are on to something? To construct meaning effectively we also need to persist over time and check our initial hunches or insights.

One implication of these observations is that it is not correct to describe the items of the *Progressive Matrices* as "problems to be solved". It is true that, once one has reached the idea behind them, one can see them as logical problems to be solved. But that is a second stage – like solving the more specific problems which emerge after one has understood something about the field of forces depicted in Figure 1.7. At that point, one may stumble on the idea of harnessing the external forces which influence the success of one's business in a manner analogous to the way in which one can harness the (invisible) equal and opposite reactions of the sea to the wind by adding a keel to one's sailing boat and thus inventing a way of driving one's boat *into* the wind instead of allowing the wind to crash it against the rocks. But who, in a bye-gone age, in their right mind have even entertained the idea that it might be *possible* to sail

a boat *into* the wind? No. It does not become a "problem to be solved" until one has stumbled on Newton's Laws and realised that, by providing the germ of a solution, they render the unproblematic problematic! How to harness social forces in an analogous way then becomes an (equally difficult) "problem to be solved" … but such thinking can only emerge as a problem *after* one has in some way "seen" – made sense of – the external field of forces.

Note that what we have said about the role of feelings, actual or mental "experimentation", and persistence in "thinking" implies that what is typically described as "cognitive" activity is primarily affective and conative – the word "conation" being a technical term for the active, as distinct from the affective, part of striving and involving will, determination, and persistence.

And here is the dilemma – for if "cognitive activity" is a difficult and demanding activity having multiple components, no one will engage in it unless they are strongly intrinsically motivated to carry out the actions which require it.

Many people do not *want* to be managers and struggle to make sense of those external economic, social, political, and cultural processes that so much determine the success of an organisation and work out how they can be harnessed or influenced. They do not *want* to think about those complex subordinates and their motives and potential talents and how these can be developed, released, and harnessed.

It is all very well to argue that, just because someone does not *want* to be a manager they will not require this difficult and demanding "ability". But what distinguishes a more from a less effective secretary? A more from a less effective machine operative? A more from a less effective sales person? A more from a less effective teacher? A more from a less effective hunter? A more from a less effective housewife?

Part of the answer is that they are more likely to think about the context in which they work and then take the initiative to improve things[1.1]. In other words, individual performance in a wide range of jobs and activities depends in part on the concerns and abilities the *Matrices* set out to measure[1.2].

Unfortunately, what we have said makes our task as psychometricians *more* rather than less difficult … because it raises the question of whether meaning-making ability can be meaningfully assessed without first finding out what people want, or tend, to think *about*. As we have said many people holding managerial positions do not want to make sense of what

subordinates have to say or wish to devise means of using the information they possess. In a sense, they are not interested in activities which would promote the survival and development of the organisation. So the organisation crashes[1.3].

But, then again, how to do something about a salesperson's observations that the product does not suit the customers, that the internal mail system loses the orders, or that the invoicing system issues incorrect invoices, loses stock, and makes problems for customers?

As Kanter[1.4] shows, taking appropriate action on the basis of such observations requires a network of people, some of whom publicise the problem, some of whom develop prototypes, some of whom find other people in other organisations who have been thinking about related issues, some of whom raise funds from government agencies, and some of whom soothe out conflicts between people who have very different motivational predispositions – but all of whom are essential to the functioning of the "group" or network.

In short, doing something about our salesperson's (or lavatory attendant's) observations requires network-based activity *around* the problem. This activity calls on talents that are rarely recognised or discussed in text books on human resource management – let alone measurable using the psychometric tools currently available to us – but all of which demand the ability to make sense of confusion and act on the insights so gained. (Kanter refers to this collection of activities as "parallel organisation" activities because they go on *in parallel with* the day-to-day operations of selling or cleaning; they do *not* **replace** them as is sometimes suggested in connection with network working. On the contrary, the selling or cleaning activities are crucial stimuli to making the observations that need to be enacted to improve the functioning of the organisation.)

So, even if someone does not want to be a manager, they are still in double jeopardy if they think you can get away without thinking. They are in jeopardy as a salesperson, for example. But they are also are in jeopardy for not contributing in their unique and indispensable way to the "parallel organisation" activity that has to take place around their job – whether that be as a salesperson, a typist, cloakroom attendant.

Yet they cannot avoid the problem by packing up and going home. For the same components of competence are required to be one or other type of effective wife, husband, lover, collaborator, friend, or political activist.

While such observations underline the pervasive importance of eductive ability, they also bring us face to face with a fundamental conceptual and measurement problem. They raise the question of whether effective panel beaters, football players, and musicians all *think* – set about trying to generate meaning –"in the same way" in these very different areas. Certainly they rarely think in words.

So, at least at the present time, it would appear that, while they are clearly onto something important, it is misleading for people like Gardner[1.5] to speak of different kinds of intelligence. It seems that the components of competence required to make meaning out of the booming, buzzing, confusion in very different areas are likely to be similar. But they will only be developed, deployed, and revealed when people are undertaking these difficult and demanding, cognitive, affective, and conative activities in the service of activities they are strongly motivated to carry out and thus not in relation to any single test – even the RPM!

As Spearman[1.6] remarked long ago, the first question is not so much "How well can someone think?" as "What does he or she tend to think *about*?" And the second is "Which of the components of effective thinking do they deploy and which do they neglect?"

Before leaving this preliminary discussion, it is convenient to make explicit a couple of other points which have hovered on the fringe of our discussion. One is that, contrary to what many highly verbal academics tend to assume, thinking is not usually verbal[1.7]. Another is that it is centrally dependent on the use of feelings and on *action* – on "experimental interactions with the environment"[1.8] designed to test the evolving understanding of the nature of "the (self-defined) problem" and the strategies required to do something about it.

What, then, do the *Raven Progressive Matrices* tests measure?

They measure what many researchers have called "general cognitive ability" – although this term is misleading because what the RPM really measure is a specific kind of "meaning making" ability. Spearman coined the term *eductive* ability to capture what he had in mind, deriving the word "eductive" from the Latin root *educere* which means "to draw out from rudimentary experience". Thus, in this context it means "to construct meaning out of confusion".

It is, however, important to note that Spearman elsewhere[1.9] noted that the range of tests from which his ***g*** – and with it "eductive" ability – had emerged was so narrow that one would not be justified in generalising the concept in the way that many authors do. There could well be other

kinds of meaning making ability that would not show up on the tests that were then available … or even constructable within current psychometric frameworks.

He made the point as follows:

> "Every normal man, woman, and child is … a genius at something … It remains to discover at what … This must be a most difficult matter, owing to the very fact that it occurs in only a minute proportion of all possible abilities. It certainly cannot be detected by any of the testing procedures at present in current usage."

We will return to the limitations of the most popular frameworks for thinking about individual differences later in this chapter and again, more fully, in subsequent chapters. But, first, how to substantiate our claim that the RPM measures at least one very important kind of meaning-making ability?

This is a much more difficult matter than many of those who have written textbooks on the subject tend to assume. As Messick[1.10] has, perhaps more than anyone else, been at pains to point out, the conceptual validity of a measure cannot in fact be established via a table of first-order correlations between various measures and criteria.

Although it may at first sight strike the reader as a strange proposition, the first step toward substantiating our claims regarding the RPM involves examining the test's conformity to the measurement model used in its development.

The Measurement Model

First let us consider what one would have to do to develop a scale to measure, or index, the "hardness" of geological substances … at the point at which one was not even sure that the concept of "hardness" had any scientific meaning.

One would first assemble a range of substances that might form a suitable set of standards against which to compare the hardness of other substances and, in this way, index, or assess, the hardness of those other substances. To this end one might assemble a range of potential reference materials – such as cotton wool, putty, cheese, PVC, plastic laminate, steel, diamond and so on.

Then one would have to show that the order was consistent – that it did not vary with such things as ambient temperature, the maturity of

the substances, or their source – and, ideally, that the differences between them were in some sense equal: that one did not, for example, have a whole lot of substances of similar, moderate, hardness and few, widely spaced, very soft or very hard ones .

To do this, one would have to show that, as one compared the substances one had chosen as candidates having for one's index with samples of all other substances, one got consistent relationships. That is, one would have to show that whenever one compared other substances with the standards one had chosen, one seldom found instances in which substances which were at some times categorised as being softer than substance number 6 in one's scale were at other times said to be harder than substance number 7.

Ideally, the word "seldom" in the previous sentence would read "never", but all measures are subject to error.

One would then discover that, for example, cheese was not a good substance to include in one's set of reference substances since its hardness would vary with source, with temperature, and with maturity. One would discard it. Hopefully one would be left with a set of substances against which it would be possible consistently to compare the hardness of all other substances.

A rather more sophisticated version of exactly this procedure was used to develop the *Progressive Matrices* tests.

Stated in the simplest possible terms, the objective was to create a set of items whose level of difficulty would increase in such a way that everyone would get all the items up to the most difficult they could solve right and fail to solve all the more difficult items. This would be the exact equivalent of a meter stick or tape measure where everyone passes every centimetre mark up to that which indicates their height and then fails to reach all the subsequent marks.

But note two things. First, at this point, it was not known whether eductive ability "exists" in the sense in which height "exists". (A better analogy is "force" because, actually, although its existence is now obvious, no such concept, let alone its measurability, existed before Newton. There was just the wind and the waves and moving planets and the Gods.)

Second, it was virtually certain that it would not be possible to find a "perfect" set of items, equivalent to the centimetre marks on a tape measure. One would have to take he average of several items to generate a reasonable index of someone's ability … just as one might take the average of three high jumps to index someone's "true" ability to make high jumps.

It is, in fact, easiest to illustrate the process used to calibrate the Progressive Matrices items, show that they formed a common scale, and discard unsatisfactory items (the equivalent of cheese in the above example) by reviewing the results of some research conducted much more recently and by at first pretending that the data relate to the measurement of the ability to make high jumps.

The graphs in Figure 1.8 show the relationship between people's high-jumping ability and their ability to clear the bar when it is set at different levels. Each graph relates to the bar set at a particular height and shows the proportion (or percentage) of people having each level of ability shown on the horizontal axis that are able to get over it.

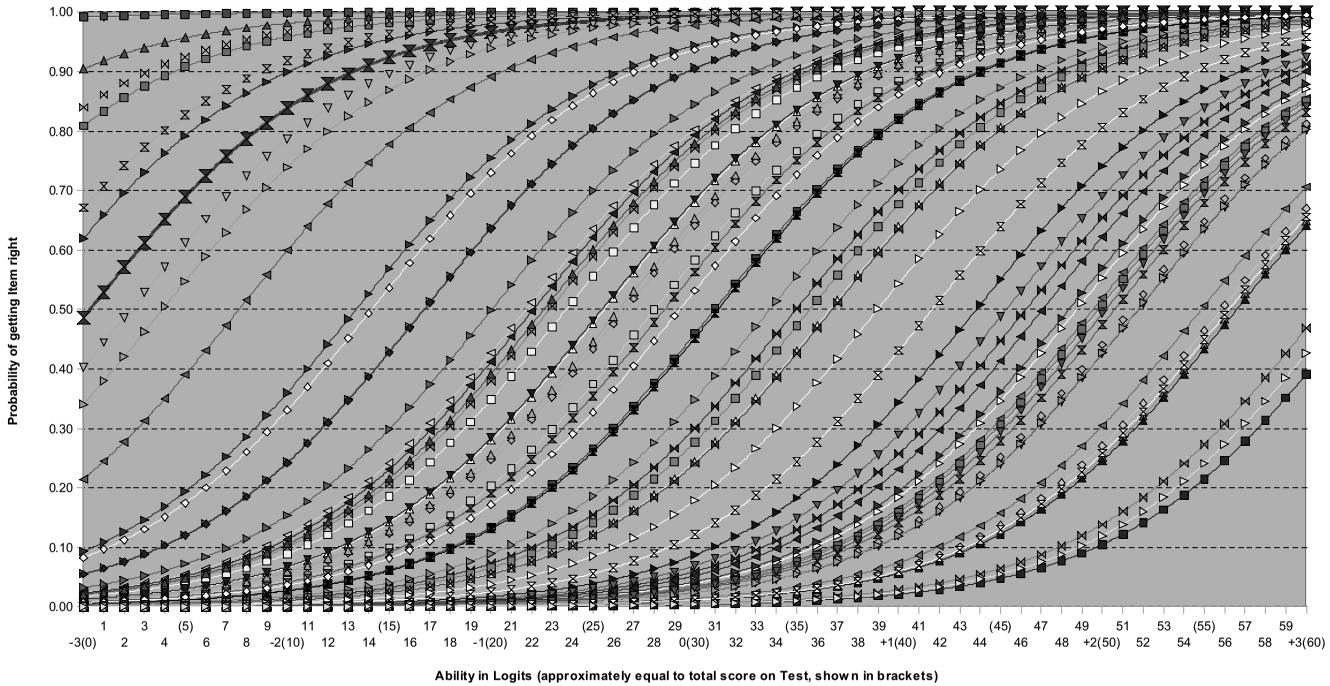
Thus, when the bar is set at very low levels – for example at the levels illustrated by the top first three curves (counting downwards) to intersect with the vertical axis – almost everyone, even of the lowest ability, is able to jump over it. But some of those with the lowest ability do knock it off. So the curves for the bar set at even the lowest levels show that only some 80 to 99% of those with the lowest ability get over it. But, of course, none of those with low ability get over the bar when it is set at high levels.

But, as we move across the Figure, we see that, at every height, the frequency with which people of somewhat similar ability get over it provides an indication of their ability.

What the overall set of curves shows is that, despite the variation in what people can do from trial to trial, one can really measure the ability to make high jumps. At all the intermediate levels at which the bar can be set, some of those who seem to have the ability to clear it knock it off and others get over it. The proportion who get over it increases from more or less zero (actually a chance level) in the Figure to almost 100% (shown in the Figures as a "probability of 1.00"). When it is set at the highest level, even the most able sometimes knock it off. The curve never reaches the 100% mark. In between, there is a consistent relationship between the curves and between them and overall level of ability. The frequency with which people of similar ability clear the bar at any one level is directly related to their overall ability. But the full range of abilities can only be assessed by changing the level at which the bar is set. Nevertheless, the curves for the bar set at these high levels conform with those obtained when the bar is set at much lower levels. They form a continuous series. They are not measuring some quite different ability.

Clearly, if we could show the same thing for the items of the RPM one would really be onto something!

Figure 1.8 *Standard Progressive Matrices* ***Plus*** Romanian Data **1-Parameter Model Item Characteristic Curves for all 60 Items**
(Each graph represents one Item)



Note: On the vertical axis, the "Probability of getting the item right" is the same thing as "Proportion of respondents getting the item right" and means the same as "Percentage of respondents getting the item right" (although the decimal point would have to be shifted two spaces to the right) and the same as "% Passes" on Figure 9.

But before we explore this possibility, let us make a make a few other observations.

First we may note that it would not make sense to set a time limit within which people have to show how high they can jump whilst also insisting that that they start by jumping over the lowest bar. Clearly, the most able would not be able to demonstrate their prowess.

Second – and this comment is mainly for the benefit of those readers who are steeped in Classical Test Theory – it would not make sense to try to establish the internal consistency (or unidimensionality) of the measure in the manner typically advocated by classical test theorists – i.e. by intercorrelating the "items" … i.e. the centimetre marks on the vertical posts … and then either factor analysing the resulting correlation matrix or calculating Alpha coefficients. This follows from the fact that, while a knowledge of whether people can clear the bar at any one level allows one to predict whether or not they will clear it when it is set at adjacent levels, it tells one very little about whether they will clear it when set very much higher. In other words, the correlations between whether or not people clear the bar at high and low levels will tend toward zero. But this does not mean that our measure of high jumping ability is meaningless[1.11]. The point can be illustrated even more strikingly by asking whether the unidimensionality (or internal consistency) of a tape measure calibrated in centimetres could be established by first giving a cross-section of people of different heights "marks" of "right" or "wrong" to indicate whether their heights were below or above each centimetre mark, then intercorrelating the "items" – i.e. the centimetre markings – across people (viz. the accuracy with which one predict from a knowledge of whether they were above or below a particular height whether they would "score" above or below each of the other centimetre marks on the tape measure), and then factor analysing the resulting correlation matrix.

A third point, related to the second, and again directed at readers steeped in classical test theory, is that, if we wish to know the correlation between ability assessed from the variance in the frequency with which people clear the bar at any one level and overall ability then the figure we need is the proportion of the variance accounted for *among those who find clearing the bar at that level problematical.* In other words, we have to exclude all those of lower and higher ability from our calculations[1.12].

Now then, as the attentive reader will already have realised from the caption on Figure 1.8, the graphs in that Figure do not in fact relate to the measurement of the ability to make high jumps but to the ability to solve

the 60 "problems" making up the most recent variant of the *Progressive Matrices* test – the *Standard Progressive Matrices* ***Plus*** (SPM+) test.

Thus, it would seem, the items of the SPM+ scale in the same way as (but perhaps not every bit as well as) the bar set at different levels when measuring the ability to make high jumps. And it also follows that it makes no sense to time the test or to seek to assess the internal consistency of the scale by intercorrelating the items (let alone factor analysing the resulting correlation matrix).

Let us now draw out a few more implications of our assertion that the value of a procedure intended to measure "meaning making ability" is to be established in exactly the same way as the quality of a scale designed to index "hardness" on the one hand or "high jumping ability" on the other.

First, the substances making up a scale to measure "hardness" (glass, steel, diamond, etc.) are qualitatively different. Yet this in no way invalidates the concept of "hardness" or its measurement. Yet the obvious qualitative differences between the items of the *Raven Progressive Matrices* has often been used to suggest that the scale as a whole lacks validity.

Likewise, no one would argue that the scalability of hardness or high-jumping ability indicates that the variance between substances or people stems from a single underlying factor. Yet many people have argued that, because the items of the RPM form an almost perfect scale, the variance in people's ability must have a single underlying cause – such as "speed of neural transmission".

Nor would they argue (as they have argued in relation to "meaning making ability") that, because, within limits, people can learn to make better high jumps, this invalidates the concept being measured.

Nor would they (as they have in relation to the RPM) set out to find single-variable explanations of the increase in high jumping ability that has occurred over the past century. Nor would they argue that, because there are no more Olympic medallists now than there were in the past, the general increase in the ability over time must be "unreal". And nor would they back-project the increases in high-jumping ability over the past century to the time of the ancient Greeks and argue that, since the Greeks were demonstrably not such poor athletes, this means that our measure of high-jumping ability must be invalid. Yet all these arguments have, in fact, been put forward to suggest that the RPM is not measuring anything "real".

At this point we have confession to make: The statistical procedures used to produce the graphs in Figure 1.8 obscure deficiencies in the test.

The test does not, in fact, perform as well as we have led you, the reader, to believe. Actually, we do not feel too bad about this deception because, as will be seen in later chapters, (a) the procedures used to produce the graphs in the Figure were not those employed in the development of the original RPM … and those graphs (see Figure 1.9) *did* reveal the deficiencies as well as the merits of the scale; (b) it was we ourselves who exposed the deficiencies in the computerised procedures used to produce the graphs in Figure 1.8 (which are also used by many other psychologists who then arrive at misleading conclusions without realising that they have done so); and (c) it is clear that, had our statistician done a better job, we could in fact have produced a test the correct graphs for which would in fact have behaved exactly as those in Figure 1.8.

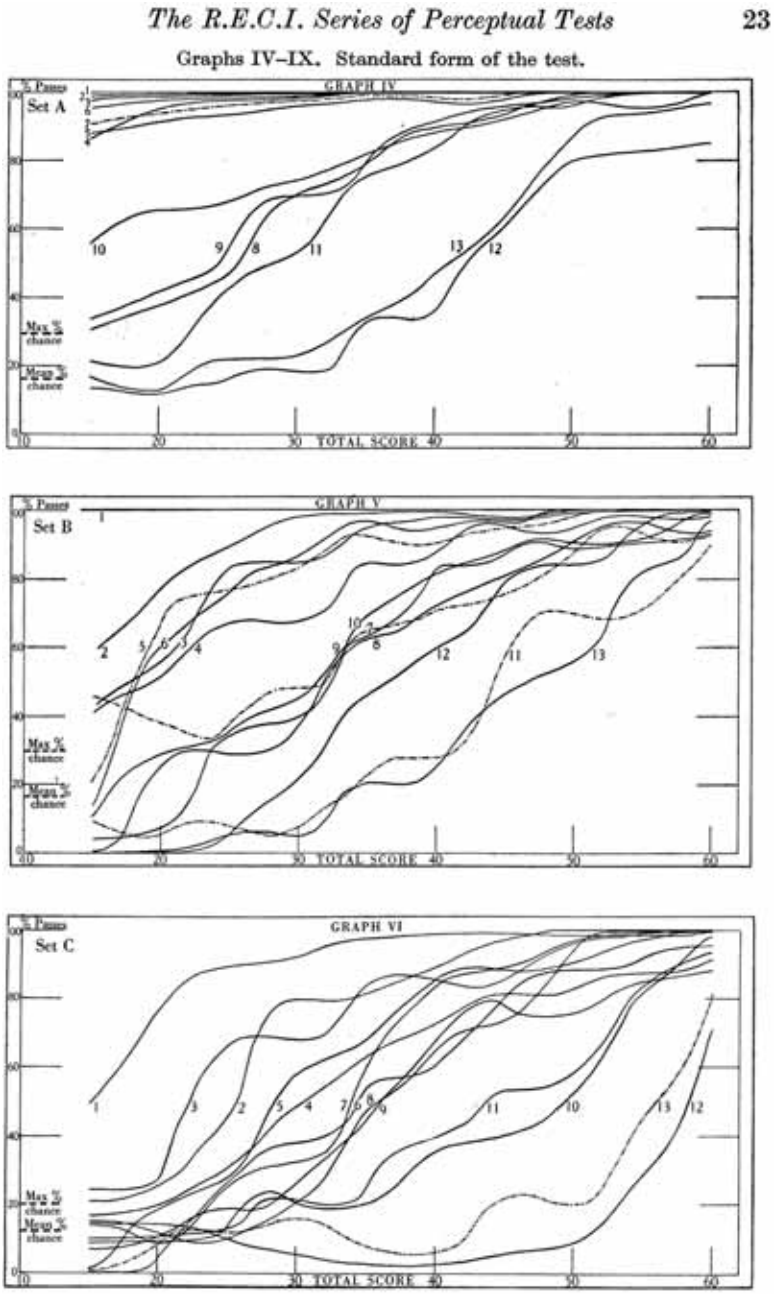
We hope that most readers will now be clear how radically the procedures used in the development of the RPM differ from those used in the construction of most other psychological tests. With the development of computers, the hand drawn graphing procedures used in the original development of the RPM have been given mathematical form, routineised, and named *Item Response Theory* (IRT), the mathematical variants of the graphs becoming known as *Item Characteristic Curves* (ICCs).

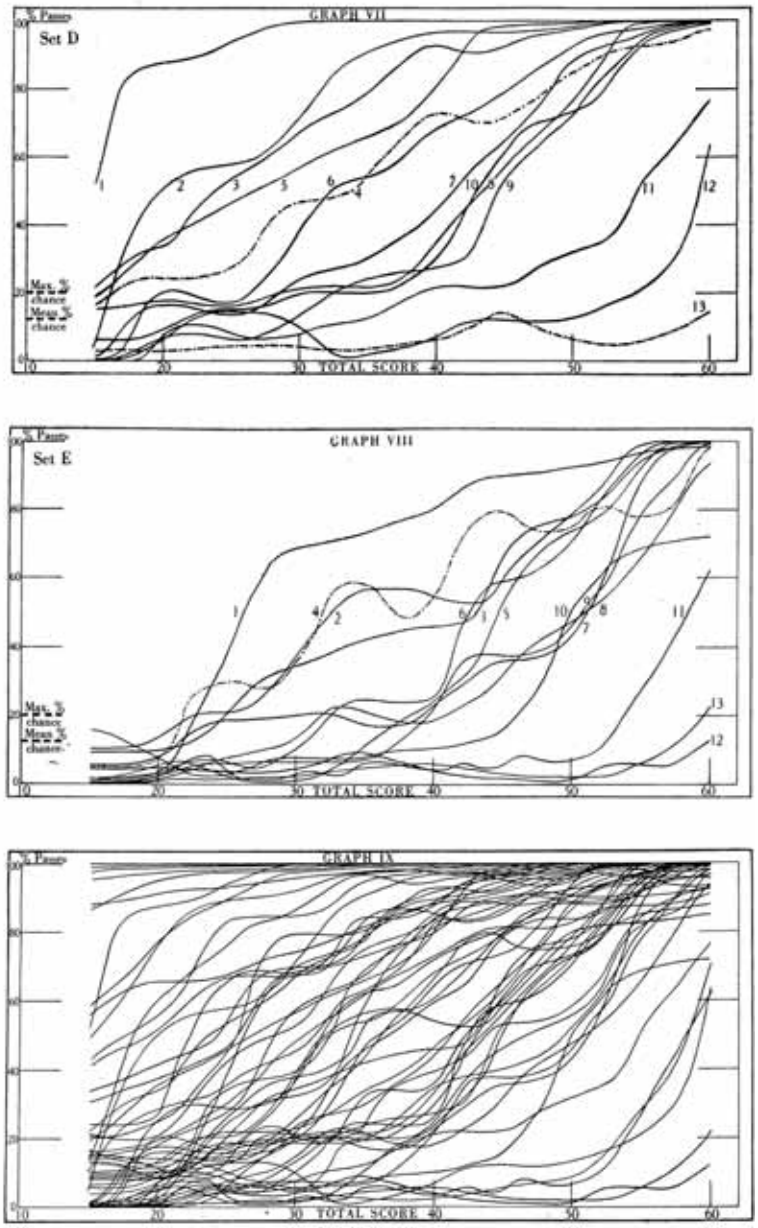
Unfortunately, because what has been said here was not so clearly articulated when the tests were first developed, few researchers understood the basis on which the tests had been constructed and this has led numerous researchers to draw misleading conclusions from their data … indeed to think in a confused manner .. and to many inappropriate applications of the tests.

The Construct Validity of the *Raven Progressive Matrices*: A Pre-Preliminary Comment!

We may now revert to our main theme: Our intention when embarking on the methodological discussion we have just completed was to begin to substantiate our claim that "eductive ability" is every bit as real and measurable as hardness or high-jumping ability. The advance of science is, above all, dependent on making the intangible explicit, visible, and measurable. One of Newton's great contributions was that he, for the first time, elucidated the concept of force, showed that it was a common component in the wind, the waves, falling apples, and the movement of the planets, and made it "visible" by making it measurable. Our hope

Figure 1.9 *Classic Standard Progressive Matrices*
Raven's Original (1939) Item Characteristic Curves





Reproduced, with the permission of the British Psychological Society, from Raven, J.C. (1939). The RECI series of perceptual tests: An experimental survey. *British Journal of Medical Psychology, XVIII*, Part 1, 16-34.

is that we have by now done the same thing for eductive, or meaning-making, ability. Our next move is to show that the conceptual framework and measurement process is much more robust and generalisable than many people might be inclined to think.

The next bit of evidence supporting this claim is striking indeed. As will be seen in more detail in a later chapter, Styles[1.13], has shown that the Item Characteristic Curves (ICCs) for a number of the tasks used to assess the Piagetian "stages of development" map directly onto the Item Characteristic Curves for the *Progressive Matrices.*

This has two important implications:

1. Taken together with our earlier observation that the obvious qualitative differences between the items in the RPM in no way undermines the case for saying that there is an underlying dimension of "general cognitive ability" (or, more correctly, "eductive" ability), they show that the early "perceptual" items form an *integral part* of the sequence which leads inexorably to the later "analytic" ones. The graphs rise hand in hand and there are no discontinuities in the sequence of items. Thus the abilities required to solve the more difficult items of the RPM are intimately associated with those required to solve the easier ones. While the abilities required to solve the more difficult ones may be layered over those required to solve the easier ones, they are not merely built upon them; they somehow integrate, or incorporate, them. Put another way, they show that "simple" perception involves the same conceptual, cognitive, affective, and conative processes as are required to make meaning out of the apparently more complex fields of thought that are so obvious in the more difficult items[1.14].
2. There are no "metamorphoses" in mental development. The apparent leaps in development that are often described as Piagetian "stages" stem simply from employing only a small number of widely spaced items to index "cognitive" development. The "stages" grade imperceptible into each other. (This implies neither that it may not be useful to discuss qualitatively different modes of thought nor that there are no metamorphoses in individual children's development … although a much more sophisticated methodology than is commonly employed would be required to detect such transformations.)

The Robustness of the Measure

So far, we have shown that the test "works" (scales) overall and argued, with some additional supporting data, that this measurability supports our claim that we are onto something important. The next step in substantiating our claim to scientific respectability has to be to show that, just as we would require anyone proposing a measure of hardness to do, that the tests' properties are invariant – that they do not vary with such things as the age, socio-economic status, education, background, and ethnicity of the respondent.

To do this while the tests were still being developed, sets of Item Characteristic Curves (ICCs) were plotted separately for children of different ages and from different socio-economic backgrounds and also for adults from a variety of occupational groups. These analyses have since been repeated using data from many countries. The conclusion is clear and very important: The test "works" – and works in much the same way – for most people from most backgrounds in most cultures[1.15]. It is therefore not possible to explain away most of the differences in average scores that exist between people from different backgrounds by arguing that the tests are, in any general sense, "foreign to their way of thought". With certain important group and individual exceptions, some of which will be discussed in later chapters, differences between groups cannot be dismissed as "meaningless." They merit investigation and explanation.

Nevertheless, it has often been argued that the "abstract" nature of the items makes them particularly difficult for "disadvantaged" children – i.e. that the test "discriminates against them". Yet it follows from the material just reviewed that this argument can, at best, be only partially true because the test works in the same way for such children as for others – i.e., despite having much the same disadvantages, there are some children who do well on the test and children from these backgrounds do not respond erratically to the problems – they do not lack familiarity with *specific* reasoning processes.

In fact Vodegel-Matzen[1.16] has provided an important direct test of the hypothesis that the "abstract" nature of the problems disadvantaged certain children. She made all the elements of which all the *Matrices* are composed more "life-like" by replacing such things as squares and triangles by everyday things like hats, bananas, and faces. Unlike Richardson[1.17], she retained the logic required to arrive at, and check, the correct answer. What then emerged was that certain types of item did become easier for some children of *all* ability levels – not just for the lower-scoring

respondents. The rank order of both items and respondents remained virtually unchanged. In other words, constructing the items of elements that it was easier to label made it easier for many people to "see what was going on" – i.e. it reduced the level of "meaning making" ability required – but the change did not differentially benefit "the disadvantaged".

History of Test Development

The *Progressive Matrices* tests were developed by J. C. Raven because he had been working with a geneticist, Lionel Penrose, on a study of the genetic *and the environmental* origins of mental defect. This meant that adults as well as children had to be tested. Those to be tested were often illiterate and thus unable to follow written instructions. But they also had to be tested in homes, schools, and workplaces which were often noisy, thus making oral questioning difficult. Raven not only found full-length "intelligence" tests cumbersome to administer, he also found the results impossible to interpret since scores on many different abilities were composited into total scores while scores on the individual sub-tests were too unreliable to use[1.18].

J. C. Raven therefore set out to develop a test which would be easy to administer, theoretically based, and directly interpretable without the need to perform the complex calculations that are often needed to arrive at scores on latent, or underlying, "factors" or variables when other tests are used.

Raven was a student of Spearman's. It is well known that Spearman[1.19] was the first to notice the tendency of tests of what had been assumed to be separate abilities to correlate relatively highly and to suggest that the resulting pattern of intercorrelations could be largely explained by positing a single underlying factor that many people have since termed "general cognitive ability" but to which Spearman gave the name "***g***". It is important to note that Spearman deliberately avoided using the word "intelligence" to describe this factor because the word is used by different people at different times to refer to a huge range of very different things[1.20]. (As we have seen, even the term "general cognitive ability" tends to have connotations about which Spearman had severe doubts.)

It is less well known that Spearman thought of ***g*** as being made up of two very different abilities which normally work closely together. One he termed *eductive* ability (meaning making ability) and the other

reproductive ability (the ability to reproduce explicit information and learned skills). He did not claim that these were separate *factors.* Rather he argued that they were *analytically* distinguishable components of ***g***.

Spearman, like Deary and Stough[1.21] later, saw this as a matter of unscrambling different cognitive *processes,* not as a factorial task. Whereas other later workers (e.g. Cattell[1.22], Horn[1.23], and Carroll[1.24]) sought to subsume these abilities into their factorial models, Spearman deliberately avoided doing so. Thus he wrote: "To understand the respective natures of eduction and reproduction – in their trenchant contrast, in their ubiquitous co-operation and in their genetic inter-linkage – to do this would appear to be for the psychology of individual abilities the very beginning of wisdom."

In addition to developing the *Progressive Matrices* test, J. C. Raven therefore developed a vocabulary test – the *Mill Hill Vocabulary Scale* (MHV) – to assess the ability to master and recall certain types of information.

At root, the *Mill Hill Vocabulary Scale* consists of 88 words (of varying difficulty) that people are asked to define. The 88 words are arranged into two Sets. In most versions of the test, half the words are in synonym-selection format and half in open-ended format. Although widely used in the UK, this test has, for obvious reasons, been less widely used internationally. Yet this test, which can be administered in five minutes, correlates more highly with full-length "intelligence" tests than does the *Progressive Matrices*[1.25].

At this point it is important to make a connection with the "fluid" and "crystallised" "intelligence" distinction developed by Cattell[1.26] and Horn[1.27] that pervades the literature

While research (see, e.g. Snow[1.28] and Carroll[1.29] for reviews) has strongly supported the eductive/reproductive distinction originated by Spearman[1.30], Horn's own review of that literature[1.31] reveals that the fluid-crystallised terminology has misled very many researchers and practitioners. What Horn shows is, in essence, that reproductive ability is *not* a crystallised form of eductive ability. The two abilities: (1) differ at birth; (2) have different genetic origins; (3) are influenced by different aspects of the environment; (4) have different neurological correlates and locations; (5) predict different things in life; and (6) change differentially over the life cycle – i.e. with age and date of birth.

The case for purging both the word "intelligence" and the fluid/crystallised formulation of the eductive-reproductive distinction from our professional vocabulary therefore seems overwhelming.

Construct Validity: Another Preliminary Statement

Having illustrated the kinds of ability the RPM and MHV were intended to measure, many readers will expect that our next step will be to review evidence bearing on the question of whether they do in fact do what they set out to do, i.e. to review research demonstrating the construct validity of the tests. Unfortunately, this turns to be much more problematic than the authors of most text books on the subject would have us believe. Because we have devoted a whole chapter of this book to showing why it is so difficult to establish the validity of a test in the classical way we will therefore, like J. C. Raven himself, duck the question for the time being and review what has emerged from some studies in which the tests have been used.

But before doing even that it is necessary to say something about the forms of the test.

Versions of the *Progressive Matrices* Tests

There are, in fact, three basic versions of the *Raven Progressive Matrices* tests, although the total comes to eight if all currently available versions are counted.

The most basic test, designed to cover all levels of ability from early childhood through adulthood to old age (and thus to facilitate research into the variations, development, and decline of eductive ability without providing detailed discrimination within any of the groups), is the *Standard Progressive Matrices.* It consists of 60 problems presented in five sets of 12. Within each Set the items become more difficult but they then revert to being easy again at the beginning of the next Set. The reason for the cyclical presentation is to provide training in the method of thought required to solve the problems … and thus to ameliorate the effects of test sophistication while at the same time providing a measure of the ability to learn from experience. This version of the test, which should be untimed, has been widely used in most countries of the world for more than 70 years. An impressive data pool suitable for cross-cultural and cross-sectional analysis has therefore been accumulated.

In order to spread the scores of the less able, a derivative of the above, consisting of the first two Sets of the *Standard* Progressive Matrices, but with a third set of easy items interposed between them was developed.

This is known as the *Coloured Progressive Matrices* since the items in the first two Sets are presented in colour.

To spread the scores of the more able, another derivative was prepared. This is known as the *Advanced Progressive Matrices* and consists of two Sets, one being a practice set of 12 items (which those who are to be tested can take home to familiarise themselves with before taking the main test – which itself consists of 36 items).

As will shortly be seen, scores on all the *Raven Progressive Matrices* tests have, in all cultures for which data exist, unexpectedly increased dramatically over the years. By the late 1980s this increase had meant that there was a marked ceiling effect among young adults on the SPM, while the APM was yielding an excellent distribution across the entire adult population. Accordingly, work was put in hand to develop versions of the tests which would (a) parallel the existing tests, both on an item-by-item and total score basis (so that all the existing normative and research data would remain applicable), and (b) restore to the SPM the discriminative power at the upper levels of ability that it had when it was first developed. This test is known as the SPM ***Plus.***

Some Findings

Heritability and the environment

Although it may seem odd to begin our review of some of the key findings emerging from research with the RPM by plunging into the contentious and difficult question of heritability, it is, in reality, important to do so because the very concept of "intelligence" is widely and inextricably bound up with assumptions about its heritability. Many researchers, such as Sir Cyril Burt, have *defined* "intelligence" as "inherited general cognitive ability". Even Flynn (who has done most to substantiate and publicise the increase in scores over time) has been inclined to argue that if, as he shows, the scores are markedly influenced by environmental variables the tests cannot really be measuring "intelligence".

Exactly the opposite position was taken by J. C. Raven. As he saw it, the first task had to be to develop a test which was theoretically based, directly interpretable, and easily administered to a cross-section of the population of all ages and coming from all socio-economic backgrounds. The last of these requirements meant that it had to be easily administered in homes, schools, laboratories, hospitals, and workplaces to people who

were often illiterate and short of time. The results obtained with such a test – and only such a test – could then be used to assess the relative impact of genetics and the environment and, most importantly, to discover *which aspects* of the environment influenced the ability being measured.

The words *which aspects* in the above sentence cannot be underlined too strongly. It is *always* possible to influence the expression of a genetic characteristic. The *only* question is *which aspects* of the environment are relevant.

It is easiest to illustrate this point by an analogy. If one takes a variety of different strains of wheat – each having, among other things, different average heights and yields – into a different environment everything changes. The average height of each strain changes, but their average heights do not remain in the same order. The average yield per hectare also changes, but the one that was "best" in one environment is not the best in another. Furthermore the correlations between height and yield change. Yet the differences between them are still genetically determined.

Having made this point, we may return to studies of the heritability of ***g*** – and eductive ability in particular.

Over the years, a number of researchers[1.32] have reported correlations between the scores obtained on a variety of measures of "general cognitive ability" by identical and non-identical twins reared in similar and different environments. Analyses of the data collected in these studies suggest that about two thirds of the variance in ***g*** is heritable, and this figure has been confirmed in what is perhaps the largest and best conducted of these studies – the Minnesota Twin Study – which employed the RPM[1.33].

The importance of genetic background in the determination of ***g*** was strikingly highlighted in the *Scottish Longitudinal Mental Development Study*[1.34]. The study was based on a representative sample of the Scottish population. In their report, the authors list the scores obtained by *all* the children in the families involved. In family after family, the variation in scores between brothers and sisters *from the same family* came to two thirds of the (huge) variation in scores in the total population. How could this within-family variation have anything other than genetic causes?

These figures cannot, however, be interpreted to mean that the environment is *un*important. As a number of researchers[1.35] have shown, effective parents create different environments for different children and children select themselves into environments in which they obtain differential treatment and this differential treatment has dramatic differential effects on their development.

These effects are stronger for qualities like creativity, self-confidence, and the ability to communicate – qualities sadly neglected by psychologists – than they are for cognitive development. However, even in relation to cognitive development, a number of researchers[1.36] have demonstrated the importance of what Feuerstein has termed *mediated learning* – i.e. children sharing in their parents' problematising, thinking about things that are not there, resolving moral dilemmas, considering the long-term social (ethical) consequences of their actions, and thereafter taking appropriate action. (The last of these involves building up their own understanding of the way society works and their place in it and learning from the effects of their actions, and thus bears directly on our introductory observations.)

Messick[1.37] succinctly captured the point that needs to be made by saying that high heritability does not imply a lack of *mutability*. (This is exactly the point made in our earlier discussion of wheat: Changes in the environment change everything, but the differences between the strains are still genetically determined.)

Changes in Scores Over Time

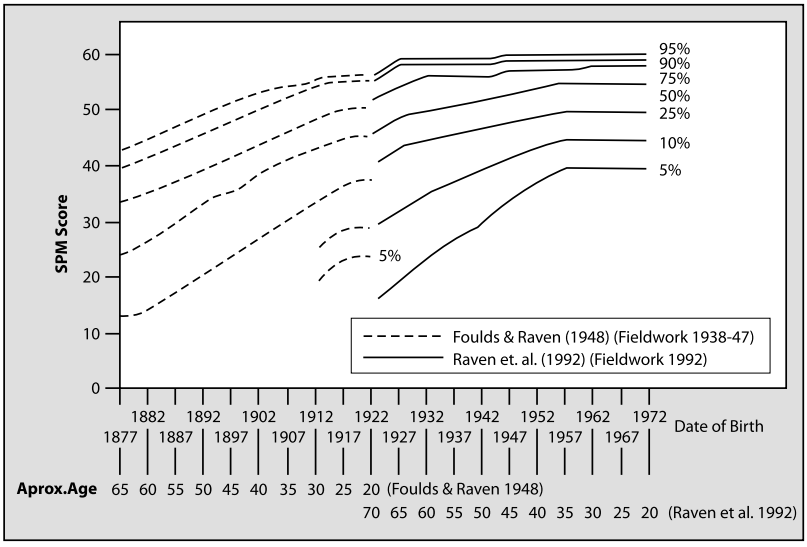
The most striking demonstration of the truth of Messick's statement so far as the RPM is concerned is to be found in research documenting huge inter-generational increases in scores[1.38].

Figure 1.10 summarises some research which will be discussed more fully in a later chapter. It shows how scores on the *Standard Progressive Matrices* have been increasing over the past century.

The horizontal axis shows both the date of birth and age of the respondents at the time of testing. Two separate samples of the adult population of Great Britain were tested circa 1942 and in 1992. The graphs in the Figure show the scores obtained by the bottom 5% of the population, the 10th percentile, 25th percentile, 50th percentile, 75th percentile, 90th percentile and the top 5% of the population[1.39] in each birth cohort. It will be seen from the Figure that scores on the RPM have increased dramatically from birth cohort to birth cohort. Since the samples for both the 1942 and 1992 studies were drawn from the same gene pool the increase could not have been produced by some genetic mechanism, but must have resulted from some environmental change[1.40].

Many researchers looking at similar data expressed in terms of means and standard deviations (but without graphing them) have concluded that

Figure 1.10 *Classic Standard Progressive Matrices*
100 Years of Eductive Ability



Note: The figure graphs the percentile norms obtained by adults of different ages (and thus birth dates) on the *Standard Progressive Matrices* when a sample was tested circa 1942 in one case and in 1992 in the other. The approximate age of people born in different years in the two samples is shown below. It will be seen that that those born in 1922 and tested circa 1942 (approximately 20 years of age when tested) obtained similar scores to those born in 1922 and tested in 1992 (when 70 years of age).

it has been the scores of the less able that have gone up most – often inferring that the increase over time has arisen from rectification of deficits in the environments of the less able. Such a view is, however, untenable. Although it would seem to be supported by the data presented in Figure 1.10, from which it will be seen that the scores of those born more recently are more bunched together than those born earlier, the bunching arises from a ceiling effect on the *Standard Progressive Matrices,* which has only 60 items. When data collected with the *Advanced Progressive Matrices* (APM), which was developed to discriminate among more able respondents, are included in the analysis, it becomes clear that the scores of the more able have also been increasing dramatically[1.41]. Just as the whole distribution of height (an equally heritable characteristic) has been moving up dramatically over the years (tall people have got still taller), the whole distribution of eductive ability scores has moved up.

In short, what these data – together with earlier data published by such authors as Bouvier[1.42], Thorndike[1.43], Raven[1.44], and Flynn[1.45] – reveal is a dramatic, and previously unsuspected, effect *of the environment* on eductive ability.

Thorndike proffered a number of possible explanations of the increase, such as changes in educational practices, increased access to television, changes in child rearing practices perhaps associated with changes in family sizes, and general "test sophistication". Most of these possible explanations have since been strongly advocated by one researcher or another[1.46] but, as will be seen in more detail later, none of these variables have the widely expected effects on RPM scores. This follows from the fact that the norms obtained at any point in time in a wide range of cultures having very different educational systems, family sizes, child rearing practices, access to television, values, levels of literacy, and calligraphies tend to be very similar. Furthermore, it has been occurring on verbal as well as non-verbal measures of eductive (meaning-making, reasoning) ability[1.47], and has been greatest among very young children who have not yet started school[1.48].

There has been a huge debate about whether the increase in scores on tests of eductive ability is "real" or due simply to such things as "test sophistication" or "familiarity with Matrices-type items". Much of the argument stems from the use of the slippery word "intelligence". No one would claim that the parallel increases in high-jumping ability or height are "unreal". So the increase in RPM scores … even educative ability scores in general … is *real.* The question is whether it has the *general* effects that many people anticipate. And here one enters the "intelligence" and "ability" quagmire because these slippery terms are often thought to refer to qualities for which there is no scientific justification but which are in turn assumed to have widespread implications for psychological and social functioning.

It is important to draw attention to an apparently insignificant feature of Figure 1.10 that has major implications for research into the development and decline of human abilities … as well as revealing that there is, in fact, a huge amount of evidence supporting the claim that eductive and many other abilities, but not reproductive ability, have increased over time.

Look at the data for the 1922 birth cohort. This cohort was about 20 years old when they were tested around 1922 and 70 when they were tested in 1992 … i.e. 50 years later. Yet the mean and distribution of their scores was almost identical at these two time points.

A number of things follow from this.

First, the scores of this birth cohort have not declined in the way most psychologists would have expected as they got older.

Ironically, J. C. Raven had interpreted the very same data collected from a cross section of the population of different ages around 1942 that we have used to plot Figure 1.10 to mean that scores did decline with age. In other words, as shown in Figure 1.11, he had plotted the 1942 data with increasing age (as distinct from date of birth) as the X axis. "Obviously", from these data, scores decline with age! It is only when the data are plotted the other way round and the 1992 data appended that the interpretation changes.

The significance of this finding cannot be overestimated.

Not only do these data reverse the interpretation of a widely reported research finding (the "decline" in intellectual abilities with increasing age) in psychology, they also show that there is, in reality, a vast pool of data (whose quality, unlike Flynn's data on changes in RPM scores over time, has never been questioned) available to support the claim that a wide range of human abilities have increased over time.

As has been mentioned, Flynn initially sought to use the evidence he had accumulated to document a dramatic effect of the environment on test scores to discredit conclusions that had been drawn about the origins of the differences between the average scores of certain ethnic groups – such as that between Blacks and Whites in America – on virtually all psychological tests. More specifically, he argued that the backward projection of the curves shown in Figure 1.10 to the time of our grandparents or the Greeks would mean that they must have had extremely low scores. Consequently, since they could not really have been that stupid, the tests must be invalid.

These arguments precipitated huge and important debates and stimulated further research. Nevertheless, the data presented in Figure 1.12 show that most of these arguments should never have occurred.

If Flynn's logic is applied to these data, they reveal that the Greeks must have had unbelievably short lives. They also discredit most of Flynn's other arguments. For example, do the changes in life expectancy over time (which must have been environmentally induced) mean that differences in life expectancy between ethnic and socio-economic groups are meaningless (as distinct from meaningful and in need of some explanation)? Are the changes over time to be explained by reference to a single underlying variable equivalent to "familiarity with Matrices

problems" or "changes in education" – or are they a result of complex and interacting changes in society? Are the factors that are responsible for the variation in life expectancy *within* a birth cohort likely to be the same as those that have resulted in the increase *across* birth cohorts – i.e. over time? Most importantly, does the fact that life expectancy is measured using a scale which confirms perfectly to the ideals, discussed above as *Item Response Theory*, which we sought to achieve in developing the RPM imply that the genetic component in that variance must have a single biological basis equivalent to the "speed of neural processing" that is so often thought to lie behind the scalability of the RPM?

Before moving on, it is, however, important to note that Flynn embarked on his research with a view to showing that, because of the impact of the environment, the differences in mean scores between ethnic groups cannot support the discriminatory educational, employment, occupational, and social policies that are often justified by reference to them. By in this way discrediting these thoughtways and associated policies he sought to advance humane ideals[1.49]. Elsewhere[1.50], he both documented the extraordinary differences between the ways in which Chinese and Blacks in American contributed to the American way of life and showed that these could not be explained by reference to differences in general cognitive ability test scores but must be due to other individual and social characteristics typically overlooked by psychologists. In short, his argument goes, the differential contributions of different ethnic groups to society cannot be attributed to differences in their cognitive ability but must be due to other (environmental?) factors that have been overlooked. A chapter summarising his vitally important work in this area will be found toward the end of this volume.

So far as can be ascertained, despite his critique of meritocracy (summarised in a later chapter) Flynn still somehow believes that the solution to the problem he poses will come from developing better measures of "intelligence" which will enable us to run a kind of meritocracy more effectively. And here, as will also be seen later, we part company with him for, as we remarked earlier when referring to Kanter's work, what seems to us to be needed is a better framework for identifying, developing and utilising the wide range of very different talents that are available in society.

Stability in Norms Across Cultures

Before summarising data showing that the *norms* for the RPM have proved unexpectedly similar across cultures with a tradition of literacy at any point in time, we must briefly review earlier data … which were equally surprising to many people at the time … supporting the statement made above that the tests "work" – scale – in much the same way in very many cultures and socio-economic groups.

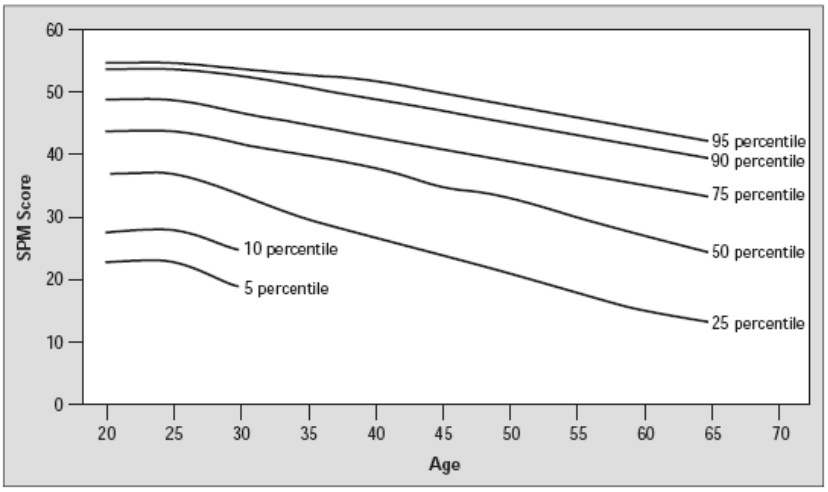
In the course of our introduction to this chapter we used graphical methods to show that the items of the RPM are not merely of steadily increasing difficulty but form a scale whereby the abilities required to solve any one item are systematically related to the abilities required to solve others (of very different difficulty) and total score. Under *Classical Test Theory*, the difficulty of an item for a particular population is indexed by the percentage of that population who get it right. Under *Item Response Theory*, it is indexed in a more sophisticated way measured in "logits". The difference between the two methods need not concern us here. What it is important to remember is the idea that the difficulty of items can be expressed in terms of mathematical indices.

These can be calculated separately for data for people from different educational, occupational, and socio-economic backgrounds as well as for people from different ethnic groups.

The correlations between the item difficulties established separately among children from eight socio-economic backgrounds (ranging from the children of professional and managerial individuals to the children of low-level manual workers such as street-sweepers) in the 1979 British standardisation[1.51] ranged from .97 to .99, with the low of .97 being a statistical rtefact. In the US standardisation[1.52], the correlations between the item difficulties established separately for different ethnic groups (Black, Anglo, Hispanic, Asian, and Navajo) ranged from .97 to 1.00. Jensen[1.53] reported similar results for the CPM. According to Owen[1.54], the test has the same psychometric properties among all ethnic groups in South Africa – that is, it scales in much the same way, has similar reliability, correlates in almost the same way with other tests, and factor analysis of these correlations yields a similar factorial structure. The correlations between the item difficulties established separately in the UK, US, East and West Germany, New Zealand, and Chinese standardisations range from .98 to 1.00.

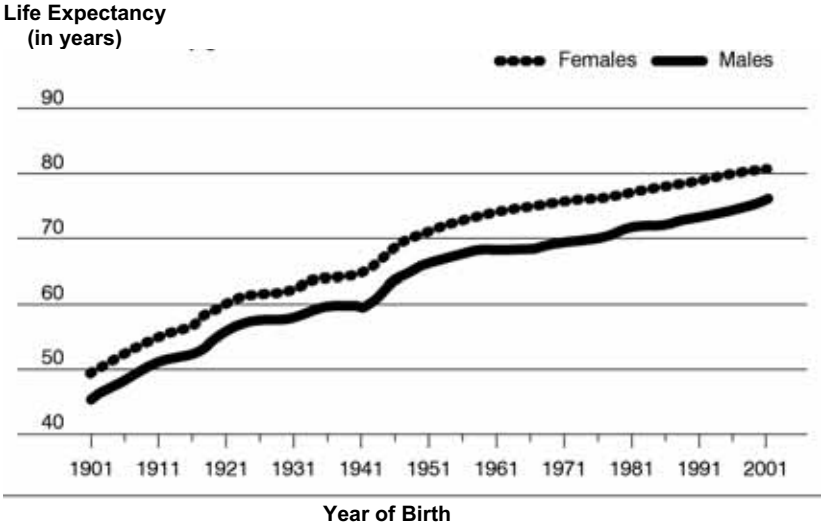
These data clearly support our earlier claim that the tests work in the same way – measure the same thing – in a wide range of cultural, socio-

Figure 1.11 *Classic Standard Progressive Matrices*
The Apparent Decline of Eductive Ability with Age
1942 Cross-Sectional Study



Note: A typical figure showing the apparent decline in *Standard Progressive Matrices* scores with increasing age among people of different levels of ability. The data was accumulated in the course of studies conducted between 1939 and 1947.

Figure 1.12 **Life Expectancy UK: Years from Birth by Gender**



economic, and ethnic groups despite the (sometimes huge) variation in mean scores between these groups.

Cross-cultural similarity in norms.

Having briefly summarised these remarkable data, we may now turn to Table 1.1 which presents a selection of cross-cultural normative data. (Readers unfamiliar with age norms presented as percentiles will find a brief explanation in Note 55 where our reasons for not presenting data in terms of Means, Standard Deviations, or Deviation IQs will also be found.)

To facilitate comprehension, many relevant columns and rows of data have been omitted from Table 1.1. Firstly, the data for very many countries for which we do have good statistics (such as Germany, France, Spain, Slovakia, Russia, New Zealand, and Australia but we have not included it here because it adds little to the observations that can be made from the data that are included. The countries that remain include several which many people would have expected to differ markedly in average ability.

Secondly, all rows of figures except those for the 5^{th} , 50^{th}, and 95^{th} percentiles have been deleted.

Thirdly, and more confusingly, the countries that are represented vary with age group. This is for no other reason than the fact that we do not have data for the full range of age groups for all the countries whose results are shown in the Table. We have therefore selected for the Table age groups for which norms for a fairly wide range of countries are available. Thus, at 10 years of age, we have included norms for the UK, USA, People's Republic of China, Pune and Mumbai (India), Qatar, Poland, and Taiwan. At 20 years of age we show the available data for the UK, Tunisia, and Belgium.

If one looks as the age groups for which data from a more complete range of countries are available – such as the 10-11 year old age group – one is first struck by the similarity in the normative data obtained from countries which have very different cultures, values, calligraphies, educational systems, access to television and computers, family sizes, religions, and child-rearing practices – and are at very different stages in "economic development". This suggests that cultural variation in these socio-demographic characteristics has much less impact that is commonly assumed.

But it is not just the similarity in the absolute level of the norms that is striking. The similarity in the variance within each of these countries is also striking. This strongly reinforces the impression that these socio-

demographic variables have relatively little effect because, if they *did* have the impact on scores that is often asserted, they would surely influence the within-culture variance. Everyone in each of these cultures is exposed to much the same cultural environment, yet it seems that it neither restricts nor enhances the within-cultural variance.

But now for an important confession. The Table does not include norms for groups which we know do not conform to this pattern: These include Blacks and Native Americans in the US (with the disconcerting exception of the Eskimos), Blacks in South Africa, Indian Tribal groups, Slovakian Gypsies, and other groups lacking a tradition of literacy. In many cases, although we know the differences exist (and are summarised in Raven, 2000 and Court and Raven, 1995), they have been established on other tests, such as the *Coloured Progressive Matrices,* and could not, therefore, have been included in Table 1.1. Nevertheless, some important recent results from substantial samples of some of these groups will be presented in later chapters of this book.

But the main point to be made here is that many cultural differences which a lot of people would have expected to have a major influence on scores appear to have, at most, a relatively minor effect.

The Occupational Predictive Validity of the RPM

Although the popularity of the RPM tests is probably based more on such things as the ease with which they can be administered to people who are unable to read or who do not speak the language of the administrator than on their demonstrated value in predicting occupational performance, their merit as the most cost-effective measure of what is generally termed "general cognitive ability" has not been unimportant.

A great deal of research conducted over many years has shown that, not only that scores on tests of "general cognitive ability" have predictive validity in the workplace, but also that the addition of data from other tests – such as of personality or specific motor skills – add little to the predictions psychologists are able to offer. Put crudely, "***g*** and not much else works". Eysenck[1.56] provided an early overview of such research in a very popular book published in 1953. There he summarised research conducted in World War II which showed that the RPM on its own was able to predict future performance as effectively as the results of full length "Assessment Centres" involving the simulation of complex real life tasks,

Table 1.1 *Classic Standard Progressive Matrices* **Some Indications of Cross-Cultural Stability** Selection of Cross-Cultural and Birth Cohort Norms Most European and Similar Norms Omitted

	Age in Years (Months)													
	8½	8½	9	9½	9½	10	10½	10½	10½	10½	10½	10½	10½	10
	8(3)	8(0)	8(9)	9(3)	9(0)	9(9)	10(3)	10(3)	10(3)	10(3)	10(5)	10(3)	10(0)	
	To	to	to	to	to		to	to	to	to	to	to	to	
	8(8)	8(11)	9(2)	9(8)	9(11)	10(2)	10(8)	10(8)	10(8)	10(8)	10(10)	10(8)	10(11)	
Percentile	UK	**KW**	UK	UK	**KW**	UK	UK38	UK	TW	PRC	PL	US	**KW**	P&M
95	42	**40**	44	46	**43**	48	48	49	52	50	49	47	**45**	46
50	**31**	**20**	**33**	**36**	**27**	**38**	33	**39**	**41**	**39**	**37**	**36**	**32**	**28**
5	13	**10**	14	14	**11**	17	13	22	23	18	13	17	**12**	11

	11	11½	11½	11½	11	12	12½	12½	12½	12	13	13½	13½	14
	10(9)	11(3)	11(3)	11(0)		11(9)	12(3)	12(0)	12(3)		12(9)	13(3)	13(0)	13(9)
	To	to	to	to		to	to	to	to		to	to	To	to
	11(2)	11(8)	11(8)	11(11)		12(2)	12(8)	12(11)	12(8)		13(2)	13(8)	13(11)	14(2)
Percentile	UK	UK	QA	**KW**	P&M	UK	UK	**KW**	US	P&M	UK	UK	**KW**	UK
95	50	51	48	**48**	49	52	53	**50**	51	52	54	54	**52**	55
50	**40**	**41**	**38**	**37**	**33**	**41**	**42**	**40**	**40**	**39**	**43**	**44**	**42**	45
5	24	25	19	**16**	12	26	27	**19**	22	14	28	29	**23**	30

	14½	14½	14	14	14	15	15½	15½	15½	15½	20	20	25	20
	14(3)	14(0)	13(9)	13(0)	13(0)	14(9)	15(3)	15(5)	15(0)	15(3)	18	18	20	18
	To	to	to	to	to	to	to	to	to	to	to	to	to	to
	14(8)	14(11)	14(2)	14(11)	14(11)	15(2)	15(8)	15(10)	15(11)	15(8)	22	22	29	22
Percentile	UK	**KW**	UK38	*AR64*	*AR00*	UK	UK	PL	**KW**	US	UK42	UK92	B	TN
95	56	**53**	53	*49*	*56*	57	57	56	**54**	56	55	59	58	56
50	**46**	**45**	44	***39***	***48***	**47**	**47**	**47**	**46**	**46**	44	**54**	**49**	**47**

AR (Argentina). The data were supplied by Lilia Rossi Case, Rosa Neer, and Susana Lopetegui. The 1964 data were collected by Direccion de Psicologia - Buenos Aires from 880 children studying in La Plata – Buenos Aires. The year 2000 data were collected by Lilia Rossi Case and her colleagues. The sample consisted of 1,740 young people who were studying, or had finished, high school or secondary level, equally distributed between males and females, plus students at public and private schools of La Plata – Buenos Aires, selected according to geographical and socio economic criteria. Full details of the study can be found in Cayssails (2001).

B (Belgium). Data collected between 1984 and 1990 by J.J. Deltour by asking students taking a course in psychometrics each to test 10 adults with equal numbers from each of four educational levels (i.e. not in such a way as to match the total population proportions from each level). The sample was neither stratified for age nor socio-economic status. See Deltour (1993).

P&M (Pune and Mumbai [Bombay], India). A carefully drawn sample of 5,161 Mumbai (Bombay) and 5,127 Pune young people were tested under the supervision of Professor C. G. Deshpande, by selected personnel from the Department of Applied Psychology, University of Mumbai and the Jnana Prabodhiai Institute of Psychology. The 78 schools involved included Government, Government Aided, and Private Schools teaching in Marathi, English, Hindi, and Gujarathi in the correct proportions. Full details are published by Manasayan (Delhi) as a Supplement to the Indian edition of the SPM Manual.

PL (Poland). Data from the 1989 Polish standardisation. See Jaworowska & Szustrowa (1991).

PRC (People's Republic of China). Data from a 1986 study of 5,108 respondents drawn from all main cities of China. Testing organised by Professor Hou Can Zhang of Beijing Normal University.

QA (Qatar). Data collected by Alanood Mubarak Ahmad AL Thani, Umm Alqura University, Saudi Arabia as part of a Masters degree programme. A random sample of 1,135 children drawn from 7 boys' and 7 girls' public elementary schools in Doha City was tested.

TN (Tunisia). Data collection organised by Riadh Ben Rejeb between 2000 and 2002 following a sampling design intended to yield 5 men and 5 women in each 5-yearly age group between 15 and 60 in each of 6 geographic areas of the country, but which, in fact, yielded a total sample of 509.

TW (Taiwan). Data collection from 2506 young people organised by Emily Miao. See Miao (1993).

UK (United Kingdom of Great Britain and Northern Ireland). Main 81/2 -15 year olds' data obtained from a nationally representative sample of UK schoolchildren, excluding those attending special schools, tested in 1979 (See Raven, J., 1981). 20 year olds' data derived from the 1992 standardisation of the SPM and APM in Dumfries, Scotland (See Raven, J., Raven, J. C., & Court, J. H., 2000). 1938 and 1942 data put together by J. C. Raven and collaborators following rocedures described in Raven, J. (2000).

US (United States of America). National norms compiled by weighting and combining a series of norms for School Districts having known demographic compositions and, as far as possible, derived from representative samples of those districts. See Raven, J. (2000).

lasting several days, and observed by a panel of professional raters. More recent summaries, covering a huge amount of data from all walks of life including the home and the community, have been provided by Schmidt and Hunter[1.57] and Gottfredson and her collaborators[1.58].

One of the most strikingly demonstrations of the inability of most other tests to add much to the predictive power of *general cognitive ability* will be found in Figure 1.13 below, which is redrawn from Jensen[1.59].

One the most impressive demonstrations of the power of *general cognitive ability* to predict social mobility (i.e. the level of job that will be attained and retained) will be found in the reports on the Scottish Longitudinal Mental Development Survey[1.60]. Using these and other data, Hope[1.61] showed: (a) that some 60% of social mobility, both upward and downward, in both Scotland and the US, can be predicted from 11 year olds' "intelligence" test scores; (b) that, by the time children are 11 years old, Scotland achieves (or did achieve) a degree of association between "intelligence" and final Socio-Economic Status (SES) that is not achieved in America until age 40; and (c) that, even when the effects of home background are partialled out, children's "intelligence" makes a major contribution to a variety of indices of their occupational success at 28 years of age. The contribution of "intelligence" is very much greater than that of educational achievement and, since the relationship does not reveal its true strength in America until 15 to 20 years after people have left the educational system, is not a surrogate for sociological tracking by the educational system.

Back to construct validity

So far so good. But the assessment of construct validity, in fact, poses a host of widely overlooked problems that we will return to in a later chapter. These include the limitations of the conceptual framework and measurement models psychologists use to think about individual differences on the one hand and the criteria of occupational performance (which, as we noted above when discussing Kanter's[1.62] work, fail to register most contributions to occupational effectiveness). Here it is more appropriate to something which just might force us to re-interpret the pattern of relationships so far discussed.

The problem is that, as Kohn and Schooler[1.63] and the author[1.64] have shown, not only do children from the same family vary almost as much in the kinds of activity they are strongly motivated to carry out (or can be said to value) as in their "intelligence", their subsequent social

mobility, both upward and downward, can be predicted every bit as well from a knowledge of the activities they are strongly motivated to carry out as from their "intelligence". People occupying different socio-economic positions vary as much in these values as in their "intelligence". Thus Kohn[1.65] showed that people occupying high socio-economic status positions in several different societies embrace activities like thinking for oneself, originality, taking responsibility for others, and initiative. In contrast, people occupying low socio-economic status positions stress toughness, strength, obedience, and having strict rules and moral codes to guide their lives. Kohn initially believed that these differences were a product of occupational experience (and, indeed, to some extent, they are). But, by sectioning the data we obtained from adolescents by origins and anticipated occupational destinations, we[1.66] were able to show that there was a great deal of variance in the concerns of children from similar backgrounds, and that this variance was related to the status of the jobs they expected to enter. This finding, like the finding that two thirds of the variance in "intelligence" test scores is within-family variance, raises serious questions about its origins. A somewhat similar finding was reported by Kinsey, Pomeroy, and Martin[1.67], who found that there was huge variation in the sexual behaviour and attitudes of children who came from similar backgrounds and that this variation predicted where those children would end up. They joined others who thought and behaved similarly. Children could hardly have learned sexual attitudes and behaviours so different from those of their parents by modelling or formal instruction. So, where does the variance come from and how does it come about that personal attitudes and behaviour of the kind exemplified by sexual behaviour come to correspond to those of the socio-economic groups people eventually enter? The variance between children from the same family has often been attributed to genetic factors, and, in this context, it is of interest that the research of the Minnesota Twin Study mentioned earlier has shown that many values and beliefs – including religious beliefs – are as heritable as "intelligence". But, if these attitudes and behaviours are not learned at work and in society, how does it come about that, in the end, children's' attitudes and behaviours tend to be characteristic of the groups with whom they end up living and working?

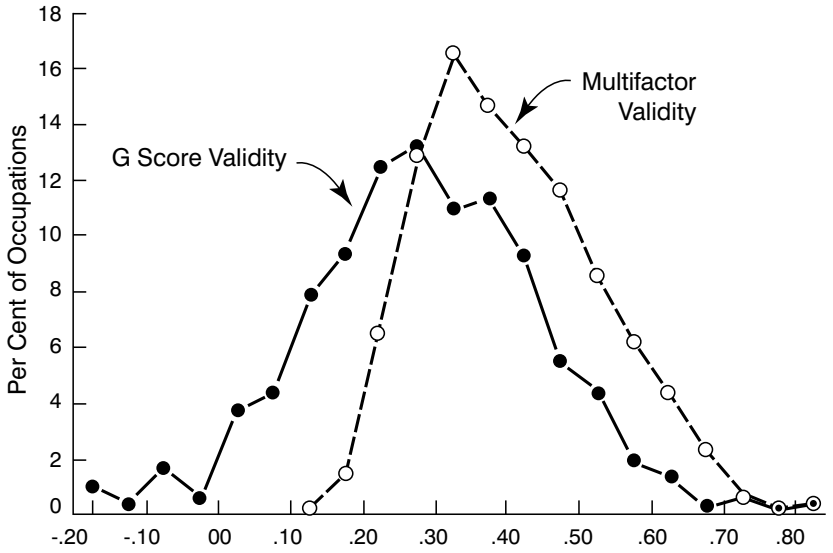
Note the problems which these observations pose for the validation and interpretation of "intelligence" tests: We have seen that children from similar backgrounds, including members of the same family, vary enormously both in their motives and values and their "intelligence". The variance in their motives predicts their future position in society every

bit as well as does their "intelligence". Which is the more basic set of variables? How does variance in "intelligence" come to be linked to variation in motives, values, and personal behaviour?

We do not, at present, know whether the portion of the variance in social position and performance that can be predicted from "intelligence" is the same as that which can be predicted from motivation and values, or whether the two are additive. So one clearly has the option of concluding that we should be focusing on the variance in the kinds of behaviour to which people are attracted and their ability to undertake those behaviours effectively rather than on their "intelligence".

Actually, this is but the tip of an iceberg to which we will return in later chapters. For, as we have seen, "cognitive" activity is a difficult and demanding activity that is heavily dependent on its affective and conative

Figure 1.13 **Predictive Validity of General Cognitive Ability** In the Context of Maximum Validities Obtainable Using All Sub-Scores of GATB



Note: Frequency distribution of 537 validity coefficients for the General Aptitude Test Battery (GATB) for 446 different occupations. The g-score is a single measure of general mental ability; multifactor validity is based on an optimally weighted composite of nine GATB aptitudes (including g) for each job category. The median validity coefficients are +.27 for g and +.36 for the multifactor composite. If g is partialled out (i.e. removed), the validity coefficient (multiple R) of the residual multifactor composite is +.24. Based on Figure 8.8, p.349, Jensen (1980). ©1980 Arthur R. Jensen. Reprinted with the permission of Arthur Jensen.

components. It follows that people cannot be expected to display their cognitive ability except in relation to activities they are intrinsically[1.68] motivated to carry out. Yet, as we have just seen, the activities people are strongly intrinsically motivated to carry out are legion and few of them have to do with generating the kinds of meaning-making ability the RPM is designed to assess. In other words people who are strongly motivated "think" about how to craft metal sheets into wonderful shapes or how to put drunks at ease or about the invisible contributions their colleagues have made to group processes are unlikely to display their abilities when asked to complete the RPM. In other words, the RPM only measures meaning-making ability *in relation to a particular kind of valued activity.* As a result, one can have little faith in generalisations about "cognitive ability" that are derived from research with the RPM. One has always to add "cognitive ability *in relation to what?"* As Spearman noted almost a century ago, the question is not usually "How *well* can they think?" but "What do they tend to think *about*?" Thinking is non verbal and emotive.

Note a very important implication of these observations: An enormous literature has grown up around the issue of the neurological localisation of "cognitive ability". Few have noted the logical error. But, as the neuropsychologist Sperry[1.69] noted, what is neurologically localised is, not "cognitive ability" but the emotional predisposition to think about particular kinds of thing.

More Recent Applications: Derived Scores

So far, we have written as if the simple raw score on the RPM is the most useful information the tests can generate. And, by and large and in the main, this is indeed the case.

But in recent years a range of new applications has emerged. While many of these (such as the assessment of "learning potential") have fallen by the wayside as a result of inadequate conceptualisation and defective measurement methodology (see the chapter on the measurement of change) the solution to the measurement of change problem appears to have opened up a range of applications associated with such things as differential reactions to drugs, stress, therapy etc. on the one hand and the detection of faked low scores in legal disputes on the other.

Notes

1.1 Kanter (1985); Schon (1973, 1983); Spencer & Spencer (1993); Raven (1997); Raven & Stephenson (2001)

1.2 Perhaps the most thorough summary of the huge amount of research demonstrating this will be found in Gottfredson (1997a).

1.3 Hogan (1990, 1991); Kanter (1985); Spencer & Spencer (1993)

1.4 Kanter (1985)

1.5 Gardner (1991)

1.6 Spearman (1926)

1.7 Despite Spearman having signalled this fact in his "*Principles of Noegenesis*" (mind creation) by saying that a knowledge of parts tends to evoke *instantly* a knowledge of a relationship (and vice versa) the contrary tends to be assumed. This is not the place to review the evidence. However a simple demonstration of the importance of non-verbal "thought" is to reflect that, when asked a question, one tends to know instantly all the topics that need to be covered in an answer, but it may take one 15 minutes or more to put them into words.

1.8 Dewey (1910); Schon (1983)

1.9 Spearman (1926)

1.10 Messick (1995)

1.11 A further implication of this observation is that it makes even less sense to submit the resulting matrix of item-item correlations to factor analysis with a view to establishing the dimensionality of the test.

1.12 To make the implications of points 2 and 3 for psychological measurement explicit, they mean that attempts to assess the internal consistency of IRT based tests by either intercorrelating and factor analysing items or by calculating item-total scores are likely to yield meaningless – entirely misleading – results. To reframe the point as a series of questions and injunctions: How meaningful would it be to seek to establish the unidimensionality and internal consistency of a tape measure by intercorrelating and factor analysing the centimetre marks? Now what happens if you make the same thought experiment in relation to assign the meaningfulness of measuring people's heights using the tape measure: This introduces error: sometimes they turn up in high heeled shoes. And then again in relation to measuring high jumping ability.

1.13 See Styles (1995); Styles & Andrich (1997).

1.14 It may be important here to counter the objection that factor analysts have isolated separate factors made up of "perceptual", "reasoning" and

"analytic" items. The first thing to be said is that, as has been shown in more detail elsewhere (Raven, Ritchie, & Baxter, 1971), the correlation matrix obtained by inter-correlating the items of a perfect Guttman or Rasch scale (such as a meter stick) can be fitted by neither a principal components analysis nor by any orthogonal or oblique rotation. The nature of a correlation matrix between the items of an IRT-based scale is determined by the properties of the scale. As we saw when discussing the measurement of high-jumping ability, knowledge of whether someone gets a very easy item right (or clears the bar at a very low level) does not enable one to predict whether they will get a very difficult item right (clears the bar at a high level). The correlation between very easy and very difficult items therefore tends to zero. On the other hand, items of similar difficulty are highly correlated since whether someone gets one item right or wrong is a good predictor of whether he or she will get the next most difficult item right or wrong. The correlation matrix obtained by inter-correlating the items after they have been arranged in order of difficulty thus has correlations which tend toward unity around the diagonal and approaching zero in the distal corners. Such a correlation matrix cannot be re-created by multiplying and adding loadings on any set of factors smaller in number than the original items. In less technical terms, factor analysis attempts to sort items into clusters that are highly correlated with other items in the same cluster but only slightly correlated with items in other clusters. If all the correlations are high around the diagonal and drop toward zero in the distal corners this cannot be done. If one insists on telling one's computer to do it, it comes up with a series of "power" or "difficulty" factors. These are made up of items of similar difficulty because adjacent items inter-correlate highly. (The average within-factor item-item correlation is determined by the number of factors one tells one's computer to extract.) But now comes the misinterpretation. Items of similar difficulty consist predominantly, though far from exclusively, of items of similar manifest content. In fact, the factors contain some of the more difficult items that come from the qualitatively different type which precede them in the test booklet and some of the easier items from the type which comes later in the booklet than the bulk of the items in the cluster. These "non-conforming" items are easily overlooked when naming the factor. So researchers have tended to name the factors they have extracted to reflect the dominant manifest content of the items in the cluster although they have, in reality, simply been grouped together because they are of similar difficulty. The correlations between the factors re conveniently ignored. Well, actually, despite the recommendations of the APA task force on statistical inference, most researchers never even look at the correlations … so they never even become aware of the problem!

1.15 The relevant statistics will be summarised later in this chapter.

1.16 Vodegel-Matzen (1994)

1.17 Richardson (1991)

1.18 Raven's (1936) observations on this point have been confirmed in numerous studies ranging from those of Spearman at the turn of the last century through Eysenck (1953) in the 1940s, Matarazzo (1990), and Deary (2000) in the present century.

1.19 Spearman (1926, 1927a&b)

1.20 These problems are still with us. The results of studies conducted with multi-component "intelligence" tests tend to be far from clear and to generate endless confused argument stemming from the wide range of constructions placed on the word "intelligence" (see, e.g. Flynn, 1984, 1987, 1999; the responses in the January 1997 edition of *American Psychologist* to the APA Statement on "*Intelligence: Knows and unknowns*" [Neisser et al., 1996], and various authors in Neisser, 1998).

1.21 Deary & Stough (1996)

1.22 Cattell (1963)

1.23 Horn (1968)

1.24 Carroll (1993)

1.25 Raven, J., Raven, J. C., & Court, J. H. (1998b)

1.26 Cattell (1963)

1.27 Horn (1968)

1.28 Snow (1989)

1.29 Carroll (1993)

1.30 In addition, Matarazzo (1990) demonstrated that the extraction of more than these two scores from multiple-factor "intelligence" tests is usually unjustified and Ree, Earles, & Teachout (1994) showed that the addition of specific factor scores to ***g*** estimates rarely improves the ability to predict occupational performance.

1.31 Horn (1994)

1.32 See, e.g., Bouchard & McGue (1981) and Plomin & Rende (1971) for reviews of the relevant studies.

1.33 See, Bouchard, Lykken, McGue, Segal, & Tellegen (1990).

1.34 See, Maxwell (1961).

1.35 e.g. Lykken (1999); Plomin (1989); Plomin & Daniels (1987); Scarr (1994).

1.36 e.g. Feuerstein, Klein, & Tannenbaum (1990); Raven (1980); Raven, J., Raven, J. C., & Court, J. H. (1998a); Raven & Stephenson (2001); Sigel (1985).

1.37 Messick (1989)

1.38 Raven (1981); Flynn (1987, 1999); Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004)

1.39 A fuller discussion of such percentile norms will be found in Note 55.

1.40 It is, of course, possible that the effect of the environment on eductive ability was mediated by a genetic change … but such an hypothesis would not find widespread support among geneticists!

1.41 Raven (2000b)

1.42 Bouveir (1969)

1.43 Garfinkel & Thorndike (1976); Thorndike (1977)

1.44 Raven (1981)

1.45 Flynn (1987)

1.46 A range will be found in Neisser (1998).

1.47 Bouvier (1969); Schaie (1983); Thorndike (1977)

1.48 Garfinkel & Thorndike (1976)

1.49 Flynn (1999, 2000)

1.50 Flynn (1989)

1.51 Raven (1981)

1.52 Raven (2000a)

1.53 Jensen (1974)

1.54 Owen (1992)

1.55 We usually present data to enable users to compare the RPM scores of any person tested with the scores of other people of a similar (or different) age and from similar (or different) backgrounds as percentile norms. The British norms for 6 ½ to 15 year olds are reproduced in Table 1.2 as an example.

The Table shows, for example, that 95% of 6 ½ year olds scored at, or below, 33, 90% at or below 30, and 5% got nine or less items correct. The other columns show the distribution of scores for the other age groups.

There are many things about this Table that are worth noting.

Firstly, the difference in score between the 90th and 95th percentile for most age groups is only two raw score points. Yet these two percentiles (often expressed as "IQs" of 120 and 125 respectively) are widely used as cut-off points for entry into educational program (such as "gifted education") that have marked differential implications for people's future lives and careers.

Expressing them as "IQs" of 120 and 125 creates the impression that there is a greater difference between such children than there really is. This means that people's lives and careers are being determined by such trivial things as whether or not they got two particular items in the test right or wrong. Some readers may be tempted to think that other tests have more discriminative power than the RPM. But this is rarely the case. Even in the so-called "high stakes testing" of the Scottish Leaving Certificate examination, the difference between the highest possible grade and a "fail" is only eight raw score marks (Spencer, E.,1979).

Secondly, the only percentile points for which raw scores are shown are the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles. Many test users – without looking at the reality on which they are based – regard such crude discrimination as inadequate for practical purposes. They want, for example, to discriminate between those with an IQ of 81 and 79, (and, indeed, as we shall see in a later chapter, precisely this difference can determine whether someone found guilty of murder will him- or herself live or die). We have already examined the reality lying behind a difference between the 95th and 90th percentiles, but the same observations apply to attempts to make fine discriminations anywhere in the distribution. Thus the use of *detailed* norms – i.e. norm tables making finer discriminations than those in Table 1.2 – is to be discouraged on the grounds that it encourages users to attach importance to small differences in score – and too much importance to "intelligence" in general. Neither the discriminative power of the tests currently available, their reliability, nor the explanatory power of the constructs they are designed to assess justifies such action. Worse, placing undue reliance on such scores undermines the quest for more comprehensive assessments and has the effect of absolving teachers, managers, and psychologists from responsibility for seeking ways of identifying and nurturing other talents.

Thirdly, although it is more obvious from Figure 1.11 (which presents similar data from the 1942 standardisation in graph form), at most ages, the distribution of scores above and below the 50th percentile is not symmetrical. As shown in Figure 1.10, this is particularly striking among adolescents born more recently. For this group there is a marked ceiling effect and very little discrimination above the 75th percentile. In other words, some of the within-age distributions are severely skewed and the extent of this skew varies with age and date of birth. In technical terms, the distributions are not Gaussian (which are often misleadingly – and entrappingly – described as "normal"). In fact a more detailed analysis shows that, at many ages, they are bi-modal (Raven, 1981). Consequently, it would be extremely misleading to attempt to summarise data such as those already presented or those to be presented in Table 1.1 in terms of means and standard deviations, and even more inappropriate to present them as deviation IQs with means of 100 and SDs of 15.

Table 1.2 *Classic Standard Progressive Matrices*
Smoothed British Norms for the Self-Administered or Group Test (Children)
From the 1979 Nationwide Standardisation

	Age in Years (Months)																		
	6½	7	7½	8	8½	9	9½	10	10½	11	11½	12	12½	13	13½	14	14½	15	15½
	6(3)	6(9)	7(3)	7(9)	8(3)	8(9)	9(3)	9(9)	10(3)	10(9)	11(3)	11(9)	12(3)	12(9)	13(3)	13(9)	14(3)	14(9)	15(3)
	to	to	to	to	to	to	to	to	to	to	to	to	to	to	to	to	to	to	to
Percentile	6(8)	7(2)	7(8)	8(2)	8(8)	9(2)	9(8)	10(2)	10(8)	11(2)	11(8)	12(2)	12(8)	13(2)	13(8)	14(2)	14(8)	15(2)	15(8)
95	33	34	37	40	42	44	46	48	49	50	51	52	53	54	54	55	56	57	57
90	30	32	35	38	40	42	44	46	47	48	49	50	51	52	53	54	54	55	55
75	22	26	30	33	36	38	41	42	43	44	45	46	47	49	49	50	50	51	51
50	**16**	**19**	**22**	**25**	**31**	**33**	**36**	**38**	**39**	**40**	**41**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**47**
25	13	14	15	17	22	25	28	32	33	34	36	37	38	39	41	42	42	42	42
10	10	12	12	14	16	17	19	23	27	29	31	31	32	33	35	36	36	36	36
5	9	10	11	12	13	14	15	17	22	24	25	26	27	28	29	30	33	33	33
n	112	138	148	174	153	166	198	172	194	187	164	164	174	185	180	196	189	191	171

Based on a nationally representative sample of British schoolchildren, excluding those attending special schools (see Raven, 1981 for details). Younger and less able children tested individually.

Fourthly, although what has been said should be sufficient to discourage users from making detailed assessments on any single test, a great deal of legislation is drafted as if it were test independent – e.g. along the lines of "all children with IQs over 125 shall be entitled to gifted education". The fact that the reference data may have been collected at different dates from samples drawn in different ways is the least of our worries. There is virtually no equivalence in the meaning of scores let alone the operational definition of "intelligence" itself across different tests. And even when the same test has been used, the actual figures which appear in tables like Table 1.2 expressed in "IQ" terms are heavily dependent on the assumptions made by the *statistician* who compiled them. For example, Dockrell (1989) has shown that the same person, tested on the same test, and judged against reference data drawn from the same standardisation sample may have an IQ of 47 if the statistician concerned made one set of assumptions and 60 if he or she made other assumptions. Yet decisions about people's lives and careers – even their right to live or die – may depend upon such insubstantial information.

To facilitate comprehension, it is perhaps worth extracting the main points of what has been said in this *note* and re-stating them as a series of bullet points:

1. Reporting data in terms of means and standard deviations can be seriously misleading because it conceals non-Gaussian distributions and thus such things as ceiling effects, the differential meaning of score differences above and below the mean, and differential change over time at different levels of ability.

2. Reporting results in terms of deviation IQs with means of 100 and SDs of 15 is even more misleading because it:

    a. Renders the non-Gaussian distributions of scores and the variation in those distributions with age even more invisible than reporting in terms of standardised deviation scores.

    b. Creates the illusion that tests have greater discriminative power than they have, thereby encouraging people to make finer discriminations than are justified.

    c. Strengthens belief in the concept of "Intelligence" and all that goes with it … such as unjustified assumptions about the generalisability of the concept of "ability" and "its" heritability.

A fuller discussion of the misleading use of such terms as "intelligence", "ability" and "reasoning ability will be found in Raven *et* al. (1998a) and a detailed discussion of the reasons for resisting the temptation to report results in terms of deviation IQs in Raven (2000a).

1.56 Eysenck (1953)

1.57 Schmidt & Hunter (1998)

1.58 Gottfredson (1997b)

1.59 Jensen (1998)

1.60 See MacPherson (1958); Maxwell (1961); Scottish Council for Research in Education (1933, 1949, 1953).

1.61 Hope (1984)

1.62 Kanter (1985)

1.63 Kohn & Schooler (1978)

1.64 Raven (1976, 1977)

1.65 Kohn (1977)

1.66 Raven (1976); Raven, Hannon, et al. (1975a&b)

1.67 Kinsey, Pomeroy, & Martin (1948)

1.68 Intrinsic is to be contrasted with extrinsic motivation, the latter encompassing many activities that are supposedly required for job performance.

1.69 Sperry (1983

References

Bouchard, T. J., & McGue, M. (1981). Familial studies of intelligence: A review. *Science, 212,* 1055-1059.

Bouchard, T. J., Lykken, D. T., McGue, M., Segal, N. L., & Tellegen, A. (1990). Sources of human psychological differences: The Minnesota Study. *Science, 250*, 223-228.

Bouvier, U. (1969). *Evolution des Cotes a Quelques Test.* Belgium: Centre de Recherches, Forces Armees Belges.

Carroll, J. B. (1993). *Human Cognitive Abilities*: *A Survey of Factor-Analytic Studies*. New York: Cambridge University Press.

Cattell, R. B. (1963). The personality of motivation of the researcher from measurements of contemporaries. In C. W. Taylor and F. Barron (Eds.), *Scientific Creativity*. New York: Wiley.

Deary, I. J. (2000). *Looking Down on Human Intelligence*. Oxford: Oxford University Press: Oxford Psychology Series No. 34.

Deary, I. J., & Stough, C. (1996). Intelligence and inspection time. *American Psychologist, 51(6)*, 599-608.

Dewey, J. (1910). *How We Think*. New York: D. C. Heath.

Dockrell, W. B. (1989). Extreme scores on the WISC-R. *Bulletin of the International Test Commission, 28*, April, 1-7.

Eysenck, H. J. (1953). *Uses and Abuses of Psychology*. Harmondsworth, Mddx: Penguin Books.

Feuerstein, R., Klein, P., & Tannenbaum, A. (Eds.). (1990). *Mediated Learning Experience: Theoretical, Psycho-social, and Educational Implications*. Proceedings of the First International Conference on Mediated Learning Experience. Tel Aviv: Freund.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29-51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171-191.

Flynn, J. R. (1989). Chinese Americans: Evidence that IQ tests cannot compare ethnic groups. *Bulletin of the International Test Commission, 28*, April, 8-20.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54(1)*, 5-20.

Flynn, J. R. (2000). *How to Defend Humane Ideals*. Nebraska: University of Nebraska Press.

Gardner, H. (1991). Assessment in context: The alternative to standardised testing. In B. R. Gifford, & M.C. O'Connor (Eds.), *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction*. Boston: Kluwer Publishers.

Garfinkel, R., & Thorndike, R. L. (1976). Binet item difficulty: Then and now. *Child Development, 47*, 959-965.

Gottfredson, L. S. (1997a). Why ***g*** matters: The complexity of everyday life. *Intelligence, 24*, 79-132.

Gottfredson, L. S. (Ed.) (1997b). Intelligence and social policy. *Intelligence, Whole Special Issue, 24*, 1-320.

Hogan, R. (1990). Unmasking incompetent managers. *Insight*, May 21, 42-44.

Hogan, R. (1991). *An Alternative Model of Managerial Effectiveness*. Mimeo: Tulsa, OK: Institute of Behavioral Sciences.

Hope, K. (1984). *As Others See Us: Schooling and Social Mobility in Scotland and the United States*. New York: Cambridge University Press.

Horn, J. L. (1968). Organisation of abilities and the development of intelligence. *Psychological Review, 72*, 242-259.

Horn, J. L. (1994). Theory of fluid and crystallized intelligence. In R. J. Sternberg (Ed.), *Encyclopedia of Human Intelligence* (443-451). New York: Macmillan.

Jensen, A. R. (1974). How biased are culture-loaded tests? *Genetic Psychology Monographs, 90*, 185-244.

Jensen, A. R. (1998). *The **g** Factor: The Science of Mental Ability.* Westport, CN: Praeger.

Kanter, R. M. (1985). *The Change Masters: Corporate Entrepreneurs at Work*. Hemel Hempstead: Unwin Paperbacks.

Kinsey, A. C., Pomeroy, W. B., & Martin, C. E. (1948). *Sexual Behavior in the Human Male*. Philadelphia, PA: W. B. Saunders Co.

Kohn, M. L. (1977). *Class and Conformity: A Study in Values*, (Second Edition). Chicago IL: Chicago University Press.

Kohn, M. L., & Schooler, C. (1978). The reciprocal effects of the substantive complexity of work and intellectual flexibility: A longitudinal assessment. *American Journal of Sociology, 84*, 24-52.

Lees, S. (1996). Strategic Human Resource Management in Transition Economies. *Proceedings of Conference: Human Resource Management: Strategy and Practice.* Alma Atat Management School, Alma Atat, Khazaksthan.

Lykken, D. T. (1999). *Happiness: What Studies on Twins Show Us About Nature, Nurture, and the Happiness Set-Point*. New York: Golden Books.

MacPherson, J. S. (1958). *Eleven Year Olds Grow Up*. London: University of London Press.

Matarazzo, J. D. (1990). Psychological assessment versus psychological testing. *American Psychologist, 45*, 999-1017.

Maxwell, J. N. (1961). *The Level and Trend of National Intelligence: The Contribution of the Scottish Mental Surveys*. London: University of London Press.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18(2)*, 5-11.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50(9)*, 741-749.

Neisser, U., Boodoo, G., Bouchard. T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loechlin, J. C. Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77-101.

Neisser, U. (Ed.) (1998). *The Rising Curve.* Washington, DC: American Psychological Association.

Owen, K. (1992). The suitability of Raven's Standard Progressive Matrices for various groups in South Africa. *Personality and Individual Differences, 13*, 149-159.

Plomin, R. (1989). Environment and genes. *American Psychologist, 44(2)*, 105-111.

Plomin, R., & Daniels, D. (1987). Why are children in the same family so different from one another? *Behavioral and Brain Sciences, 10*, 1-15.

Plomin, R., & Rende, R. (1971). Human behavioral genetics. *Annual Review of Psychology, 42*, 161-190.

Raven, J. (1976). *Pupil Motivation and Values*. Dublin: Irish Association for Curriculum Development.

Raven, J. (1977). *Education, Values and Society: The Objectives of Education and the Nature and Development of Competence*. London: H. K. Lewis (now available from the author at 30, Great King Street, Edinburgh EH3 6QH).

Raven, J. (1980). *Parents, Teachers and Children: An Evaluation of an Educational Home Visiting Programme*. Edinburgh: Scottish Council for Research in Education. Distributed in North America by the Ontario Institute for Studies in Education, Toronto.

Raven, J. (1981). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.1: The 1979 British Standardisation of the Standard Progressive Matrices and Mill Hill Vocabulary Scales, Together With Comparative Data From Earlier Studies in the UK, US, Canada, Germany and Ireland*. San Antonio, TX: Harcourt Assessment.

Raven, J. (1997). *Competence in Modern Society: Its Identification, Development and Release*. Unionville, New York: Royal Fireworks Press. First published in 1984 in London, England, by H. K. Lewis.

Raven, J. (2000a). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.3 (Second Edition): A Compendium of International and North American Normative and Validity Studies Together with a Review of*

the Use of the RPM in Neuropsychological Assessment by Court, Drebing, & Hughes. San Antonio, TX: Harcourt Assessment.

Raven, J. (2000b). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1-48.

Raven, J., Hannon, B., Handy, R., Benson, C., & Henry, E. A. (1975a). *A Survey of Attitudes of Post Primary Teachers and Pupils, Volume 1*: *Teachers' Perceptions of Educational Objectives and Examinations*. Dublin: Irish Association for Curriculum Development.

Raven, J., Hannon, B., Handy, R., Benson, C., & Henry, E. A. (1975b). *A Survey of Attitudes of Post Primary Teachers and Pupils, Volume 2: Pupils' Perceptions of Educational Objectives and their Reactions to School and School Subjects*. Dublin: Irish Association for Curriculum Development.

Raven, J., Raven, J. C., & Court, J. H. (1998a). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1*: *General Overview*. San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (1998b). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 5*: *The Mill Hill Vocabulary Scale*. San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3*: *The Standard Progressive Matrices, Including the Parallel and Plus Versions*. San Antonio, TX: Harcourt Assessment.

Raven, J., Ritchie, J., & Baxter, D. (1971). Factor analysis and cluster analysis: Their value and stability in social survey research. *Economic and Social Review*, 367-391.

Raven, J., & Stephenson, J. (Eds.). (2001). *Competence in the Learning Society.* New York: Peter Lang.

Raven, J. C. (1936). *Mental tests used in genetic studies*: *The performances of related individuals on tests mainly eductive and mainly reproductive.* Unpublished Master's Thesis, University of London.

Raven, J. C. (1939). The RECI series of perceptual tests: An experimental survey. *British Journal of Medical Psychology, XVIII*, Part 1, 16-34.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than ***g***. *Journal of Applied Psychology, 79*, 518-524.

Richardson, K. (1991). Reasoning with Raven – in and out of context. *British Journal of Educational Psychology, 61*, 129-138.

Scarr, S. (1994). Culture-fair and culture-free. In R. J. Sternberg (Ed.), *Encyclopedia of Human Intelligence*. New York: MacMillan.

Schaie, K. W. (Ed.). (1983). *Longitudinal Studies of Adult Psychological Development*. New York: Guilford Press.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124(2)*, 262-274.

Schon, D. (1973). *Beyond the Stable State*. London: Penguin.

Schon, D. (1983). *The Reflective Practitioner*. New York: Basic Books.

Scottish Council for Research in Education (1933). *The Intelligence of Scottish Children*. London: University of London Press.

Scottish Council for Research in Education (1949). *The Trend of Scottish Intelligence*. London: University of London Press.

Scottish Council for Research in Education (1953). *Social Implications of the 1947 Scottish Mental Survey*. London: University of London Press.

Sigel, I. E. (Ed.). (1985). *Parent Belief Systems: The Psychological Consequences for Children*. Hillside, NJ: Erlbaum.

Snow, R. E. (1989). Toward assessment of cognitive and conative structures in learning. *Educational Researcher, 18(9)*, 8-14.

Spearman, C. (1926). *Some Issues in the Theory of **g** (Including the Law of Diminishing Returns)*. Address to the British Association Section J – Psychology, Southampton, England, 1925. London: Psychological Laboratory, University College: Collected Papers.

Spearman, C. (1927a). *The Abilities of Man*. London, England: MacMillan.

Spearman, C. (1927b). *The Nature of "Intelligence" and the Principles of Cognition* (Second Edition). London, England: MacMillan.

Spencer, E. (1979). *Folio Assessments or External Examinations?* Edinburgh: Scottish Secondary Schools Examinations Board.

Spencer, L. M. (1983). *Soft Skill Competencies*. Edinburgh: Scottish Council for Research in Education.

Spencer, L. M., & Spencer, S. M. (1993). *Competence at Work*. New York: Wiley.

Sperry, R. (1983). *Science and Moral Priority: Merging Mind, Brain, and Human Values*. Oxford: Blackwell.

Styles, I. (1995*). Integrating Quantitative and Qualitative Approaches to Intelligence: The Relationship Between the Algorithms of Raven's Progressive Matrices and Piagetian Stages*. Paper presented at the Annual Conference of the American Educational Research Association, San Francisco, 1995.

Styles, I., & Andrich, D. (1997). Faire le lien entre variables psychometriques et variables cognitive-develomentales regissant le functionnement intellectual. *Psychologie et Psychometrie, 18(2/3)*, 51-78.

Thorndike, R. L. (1977). Causation of Binet IQ decrements. *Journal of Educational Measurement, 14*, 197-202.

Vodegel-Matzen, L. B. L. (1994). *Performance on Raven's Progressive Matrices*. Ph.D. Thesis, University of Amsterdam.

Chapter 2

Linking Psychometric and Cognitive-Developmental Frameworks for Thinking About Intellectual Functioning

Irene Styles*

Abstract

Two major approaches to understanding intellectual functioning – the psychometric and the cognitive-developmental approaches epitomised by the seminal work of Binet and Piaget, respectively – are here considered complementary rather than incommensurable and, in particular, as essentially manifestations of the same underlying construct but at different levels of scale. From this perspective, and by exploiting Item Response Theory, performances of persons on *Raven's Progressive Matrices* (exemplifying the psychometric approach) and performances on three Piagetian tasks (the Balance, Chemical Combinations, and Correlations tasks) are mapped onto a single continuum of intellectual development. *The implication is that qualitative and quantitative conceptions of intellectual development are closely interlinked: within each cognitive-developmental stage, a series of small, incremental, quantitative changes occur and evolve into a major qualitative change in cognitive functioning.* In order to clarify the nature of the transformations in thinking that occur at the transition points between one Piagetian stage and another new taxonomy is developed to classify RPM items. Knowledge of the RPM items which are operational at each Piagetian transition indicates the transformations in thinking that are required.

* Contributions from discussions with David Andrich are gratefully acknowledged.

This research was supported in part by the Australian Research Council and in part by a Special Grant from Murdoch University.

Helen Shurven, Mary Kepert, and Chris Batini provided excellent help in the collection of data.

The participation of two private schools in Perth, Western Australia was central to the project, as was permission from J. Raven to computerise the *Raven's Progressive Matrices*.

This paper is divided into two Parts. Part I describes the data base for both studies and outlines the way in which both the difficulty of Raven Progressive Matrices (RPM) items and developmental levels in thinking as revealed by interviews carried out using three Piagetian tasks were mapped onto a common continuum … and the conclusions to be drawn from the demonstration that it is, indeed, possible to do so. Part II develops a new taxonomy for classifying RPM items and examines those that are operational at each Piagetian transition point.

PART I

The development and expression of intellectual functioning has been, and is, of major interest to psychologists and educationalists. Although a range of different perspectives for studying intellectual development have been put forward, two major perspectives dominate: one, which has developed out of the work of Binet, is often termed the *psychometric approach*; the other, which has developed out of the work of Piaget, is often termed the *cognitive* or *stage-developmental* approach.

According to the psychometric perspective, intelligence is conceptualised as a set of quantifiable dimensions along which people can be ordered (Seigler & Richards, 1982) and along which people may advance as growth takes place. According to the cognitive or stage-developmental perspective, people pass through four major stages identified by qualitatively different modes of reasoning which are universal and irreversible and which transcend substantive domains, although the rate of development may depend on individual differences, including differences in experience (Piaget, 1972). Development through the stages is not discontinuous, but within each stage a person may reason in a relatively stable way for some time before moving relatively rapidly into the next stage (Case, 1978).

Many tasks have been developed to operationalize both these perspectives. In the psychometric approach, a set of relatively homogenous tasks relevant to a particular dimension of intelligence and of increasing difficulty, are constructed and the number of correct responses a person gives constitutes his/her location on the dimension. In the stage-developmental approach, sets of questions relating to a variety of tasks (usually involving concrete materials) are used to elicit the kind of reasoning persons employ to solve a given problem and, on the basis of the nature

of their reasoning, persons are deemed to be operating at Stage I (pre-operational), Stages IIA or IIB (concrete operational, A or B), or Stages IIIA or IIIB (formal operational A or B) (Inhelder & Piaget, 1958). The quality of reasoning, rather than the "correctness" of the solution, is the relevant information which is assessed.

These two approaches have been viewed, mostly, as mutually exclusive, although some researchers have considered them complementary (Elkind, 1974; Flavell,1971). The latter view is taken in this paper. Historically, this complementarity seems to fit well with the fact that Piaget worked with Binet and began to develop his theory as a result of focussing on the errors children tended to make when attempting the Binet items and noticing consistencies amongst those errors. Substantively, similarities between the two include a view of intelligence as having an adaptive function, as changing in some way with age, and as becoming more complex and stable over time. The common charge (e.g., Farnham-Diggory, 1972) that a psychometric approach focuses exclusively on an outcome or product whereas a cognitive-developmental approach is concerned with the processes involved in thinking about and solving a task, seems too simplistic: the assumption of particular processes required to solve test items correctly is implicit (and is often explicit) in a psychometric approach (for example, Spearman postulated eduction of relations and correlates as being the fundamental processes associated with general intelligence) and, conversely, there are outcomes in Piagetian tasks which are judged as more or less successful at particular levels relative to an expected, correct outcome. The difference seems to be one of the focus of the task presented and, therefore, the level or scale of the assessment of the processing, rather than a different perspective altogether.

Methodologically, there are also similarities between the two approaches: for instance, from a psychometric perspective, Binet developed items that were judged appropriate for particular age levels (children of a specific age were expected, in general, to be able to complete these particular items successfully) and which were ordered in difficulty, that is, the items can be conceived as being on a continuum of increasing difficulty. From a cognitive-developmental perspective, children of a particular age are expected to be able to reason at a level generally characteristic of that age. Further, the notion of stages in development implies an order, a direction and a hierarchy of difficulty of tasks, or levels within tasks, that can also be conceived as being on a continuum. "Cognitive stages have a sequential property, that is, they appear in a

fixed order of succession" (Piaget, 1970) and "each one of these periods or subperiods is necessary to the constitution of its successor" (Piaget, 1970) – any one stage is both an extension of the one before and the basis of the one following (Piaget, 1970). For these reasons, the notion of integrating psychometric and stage-developmental approaches to intellectual development is theoretically consistent.

Previous research has attempted to investigate the relationship between psychometric measures of intelligence and Piagetian measures using factor analysis. Earlier conclusions from this research were that the two measures were distinct (de Vries, 1974; de Vries & Kohlberg, 1977) and that they, therefore, addressed different types of intellectual functioning (although some of the research was questioned on the grounds of faulty methodology (Carroll, Kohlberg & de Vries, 1984)). A more recent attempt at explicating the relationship between the two approaches (Lim, 1988) concluded that they both contributed to a ***g***, or general intelligence factor and that, in addition, there was a distinct formal operational factor as well as spatial, numerical and verbal factors. However, a major weakness of factor analysis as a methodology is that the correlations reflect the relationship between the relative difficulty levels of the items and the location of the abilities of the people relative to these items rather than the structure of the variables (Duncan, 1985; Styles & Andrich, 1994). Factors may then be, at best, no more than difficulty factors.

Mathematical Modelling of Intellectual Development

Little work has been done in applying formal mathematical modelling to the study of aspects of Piaget's theory, although three exceptions have been Andrich and Constable (1984), Bond (1993), and Davison, King, Kitchener, and Parker (1980).

> "Clearly, the sequencing of stages implies an ordered and hierarchical progressing, but there are two mechanisms which may satisfy such a structure. One is the *cumulative* mechanism most simply understood as exhibiting a Guttman scale. According to this mechanism, if one can reason at a given stage, then one can reason also at earlier stages. The other is the *unfolding* mechanism, most simply understood in terms of Coombs' work in choice data. According to this mechanism, if a person is to choose among a set of entities which have been ordered according to some principle, then the person will tend to choose the ones nearest to his or her ideal point and

> tend to reject ones at greater distances from the ideal point, irrespective of direction. Analogously, a person will tend to apply a stage of reasoning corresponding to his/her present stage, and tend not to apply reasoning at either an earlier or a later stage" (Andrich & Constable,1984)

In the case of Piagetian tasks (as is the case in most psychometric tasks), the mechanism of responding to items is cumulative in that if a person is reasoning at, say, Stage IIB, then that person is deemed to have already passed through the lower stages I and IIA and not yet reached Stages IIIA or IIIB. Thus, although *development* is seen as unfolding in that a person at, say Stage IIB will not reason at lower level, the *measurement* of that development is cumulative.

In addition to articulating a mechanism which can be modelled, it is important to conceptualise the response mechanism as probabilistic, rather than deterministic. In a probabilistic model, a particular person will have a higher probability of responding in a mode appropriate to a certain stage than s/he would have of responding at a stage higher or lower, but some deviations from the expected stage (that is, the stage of highest probability) are consistent with this conceptualisation. Thus it is not expected that the manifested and observed reasoning will accord perfectly and be recorded unequivocally as being at a particular stage, as it is in a deterministic model. This probabilistic formalisation is consistent, for example, with Chomsky's (1976) conceptualisation of the difference between competence and performance, in that a number of influences (such as anxiety or a lapse of concentration) may cause a person to reason at a lower stage (level of performance) than that person is actually capable of (level of competence).

From this perspective, it should be possible (as is the case with psychometric test items) to order cognitive-developmental transition points between levels of reasoning on different tasks along a continuum of difficulty. The idea of such an ordering of cognitive tasks involving qualitative data was first formulated by Thurstone (1925). For intelligence, aptitude or achievement tests, he considered the possibility of mapping the levels of difficulty of tasks onto a continuum according to their relative difficulties.

We shall, therefore, locate these test questions on the scale as landmarks of different levels of intellectual growth (Thurstone, 1925).

In the same way, attitude statements could be characterised as landmarks on an affective continuum according to their relative affective values. Thurstone's work provides a conceptualisation for reconciling

psychometric and stage-developmental theories which was elaborated by Andrich and Constable (1984), using Rasch (1960/1980) models for measurement. In particular, one of these probabilistic mathematical models – the Extended Logistic Model (ELM) – provides a method for attaining this theoretical reconciliation by allowing the simultaneous mapping of dichotomous and polychotomous items (Andrich,1985). Dichotomous task (item) difficulties located on a continuum can be considered as single transition points or thresholds *between* tasks, whereas Piagetian tasks have more (polychotomous) responses with transition points *within* tasks. Note again that the response mechanism underlying the model is a probabilistic rather than a deterministic one. If it can be shown that the data of interest fits the model, then that data can be considered, at one level, to form a continuous, unidimensional scale. The methodology provides the possibility of investigating whether the data conform to the model and, therefore, whether the theoretical reconciliation is tenable.

Method

Research Design

A combination of longitudinal and cross-sectional designs was used to examine within and between cohort differences in intellectual development over a period of 6 years. Initially, three age cohorts were tested (10, 12, and 14 years) with roughly equal numbers of children (60) in each group and each group consisted of approximately equal numbers of boys and girls. Mean ages of the three groups at the first test occasion were 10.03, 12.09 and 14.07 years. All the children came from two Perth (Western Australia) metropolitan schools and from medium to high socio-economic status families. Children were matched on birthdate across the three groups for the two sex groups separately. Each child was tested at six monthly intervals on the psychometric variable (the RPM). After the first test occasion, a subset of 60 children was selected – equal numbers from each age and sex group – to be tested on the stage-developmental variable (three Piagetian tasks) at yearly intervals. These children were selected to be representative of the ability range of the entire sample of children as shown by performances on the psychometric variable on the first occasion.

Parental permission to participate was obtained for each child and only one child was excluded because parental permission was not given. The attrition rate was small: under 6% over six years. Children were tested a maximum of 10 times on the psychometric variable and four times on

the stage-developmental variable. They were not tested beyond age 16, hence there was a limited amount of data collected on the children who were initially 14 years old.

The Psychometric Variable – Raven's Progressive Matrices

The variable chosen to study intellectual development from a psychometric perspective was operationalised by the *Raven's Progressive Matrices* (RPM). Raven described his test as "a test of person's present capacity to form comparisons, reason by analogy and develop a logical method of thinking regardless of previously acquired information" (Raven, in Burke, 1958). It was originally developed to assess one of Spearman's (1927) two components of "**g**" or general intelligence – that of eductive ability (Raven, J. C., 1940; Raven, J., 1989), the other being reproductive ability. This has been supported by the results of Snow, Kyllonen and Marshalek's (1984) study which showed the RPM to measure abilities central to the concept of general intelligence.

The many advantages of this test, the reasons it was chosen and the way it was computerised and administered, has been described in detail in Styles and Andrich (1993).

There are four forms of the Matrices: the Coloured (suitable for persons less than 10 years of age), the Standard (for persons from about 10 to 15 years of age) and Advanced Sets I and II (for adults or able younger persons). All forms except the Coloured form were used in this study. The initial item in each of the five sets of items in the Standard form were used as examples. All the rest (103 items in all) were ordered in difficulty and administered individually in computerised format. Details of the administrative procedures can be found in Styles (1991).

The Stage-Developmental Variable – Three Piagetian Tasks

The variable chosen to define intellectual development from a stage developmental approach and, in particular, to focus on formal operational thinking, was operationalised by performance on three Piagetian tasks: Equilibrium in the balance, Chemical combinations and the Correlations task. Inhelder and Piaget (1958) have discussed the transformation of modes of thought pertinent to the emergence of formal operational thinking. These operations include conservation, combinatorial operations, the notions of inversion, reciprocity and proportionality, and correlation. In this study, tasks to provide evidence of the level of chidren's thought processes involving all but the first-mentioned operation (conservation)

were used. Piaget & Inhelder's discussion seems to indicate these operations are ordered at one level (Inhelder & Piaget, 1958), therefore it is possible to conceive of three of these operations (the proportional, combinatorial and correlational operations) as being of increasing difficulty and, in fact, requiring increasing competence, with competence in the earlier ones being necessary for competence in the last. In particular, the combinatorial operation seems to be regarded as fundamental in the development of formal operations. "The combinatorial operations do not actually belong to the set of propositional operations and do not derive from them; on the contrary, they are the prerequisite conditions of their development (Inhelder & Piaget, 1958). The proportionality schema is seen as effecting "the transition between schemata" (Inhelder & Piaget, 1958) and is "inherent in the integrated structure which seems to dominate the acquisitions specific to the level of formal operations" (Inhelder & Piaget, 1958). And, finally, correlation is related to the concept of proportions and "depends on the propositional combinatorial system (Inhelder & Piaget, 1958). Thus, all three operations appear to be crucial in the development of formal operations and it was deemed important to include tasks designed to elicit them in the study. Although all of them are necessary in attaining the level of formal operations, as is the case with all Piagetian tasks, stages of development within each task toward full use of these operations can be identified and have been characterised accordingly (Inhelder & Piaget, 1958).

The three specific tasks chosen were (1) Combinations of Coloured & Colourless Chemical Bodies (Chem); (2) Equilibration in the Balance (Bal); and (3) Correlations (Corr). Inhelder & Piaget (1958) provide a detailed description of these tasks.

Administration

All three tasks were administered in the same order (Chem, Bal and Corr) once a year for three years and the last occasion (fourth) occurred after eight months, because a group of children were about to leave school.

On occasions two to four, the Chemical Combinations task was replaced by the electrical equivalent of the task (Philp & Kelly, 1974) because the administration was quicker and less messy and these were important factors to consider when the children could spend a limited amount of time away from their classes.

Questions were presented in a semi-structured format in that a basic set of questions was presented, but the administrator used additional

probing questions when necessary. All children were given pencil and paper to use if they wished. Administrators were not aware of the level of the child's intellectual performance on the Matrices or in academic areas, except for one administrator on the first occasion who was aware of the children's general level of performance on the RPM. All interviews were recorded on tape and assessed by two judges according to coding schedules developed from the characteristics of each stage of reasoning as described in Inhelder and Piaget (1958).

Coders were trained in assessing the interviews using a few interviews which could not be used because the respondents had left the schools. The coding of the first three sets of data were checked for reliability by a second coder. The inter-judge reliability was 73%. The coding for a subset of the last set was checked by a second coder and any discrepancies arising by either method were resolved by discussion between the two coders.

Data Analysis

All responses to the RPM and the three Piagetian tasks were analysed according to the Extended Logistic Model and mapped onto the same continuum. Specifically, the location of the items of the RPM and the thresholds marking off the transition points between the stages of the cognitive developmental tasks were located on the same continuum. The procedure used to analyse the data involved a pair-wise estimation algorithm, often employed by Choppin (1968, 1983), which accounts routinely for missing data and therefore permitted a joint analysis of data where more children had responded to the RPM than to the Piagetian tasks and where not all the RPM items were attempted by everyone on all occasions. The computer program used was ASCORE (Andrich, Lyne, & Sheridan, 1990), which also provides statistical tests of fit enabling a check of the conformity between the data and the model.

Results

When the psychometric (103 RPM items) and the stage-developmental (3 Piagetian tasks) variables were scaled jointly using the unidimensional Extended Logistic Model, the fit to the model was satisfactory. These results mean that the psychometric and stage-developmental variables formed a unidimensional scale on a single latent trait at the level of

precision provided by the data. From these results, the interpretation is that, at one level, the reasoning processes underlying both types of variable are similar: the underlying latent trait is postulated to be that of abstract reasoning – considered by Elkind and others to be the fundamental reasoning process in intelligent thought (Elkind, 1974).

Because the item difficulties and person ability estimates are in the same metric, and given that the items fit the model, the difficulties of all the items relative to each other can be examined. These are depicted in Figure 2.1 along with the frequency distribution of the persons.

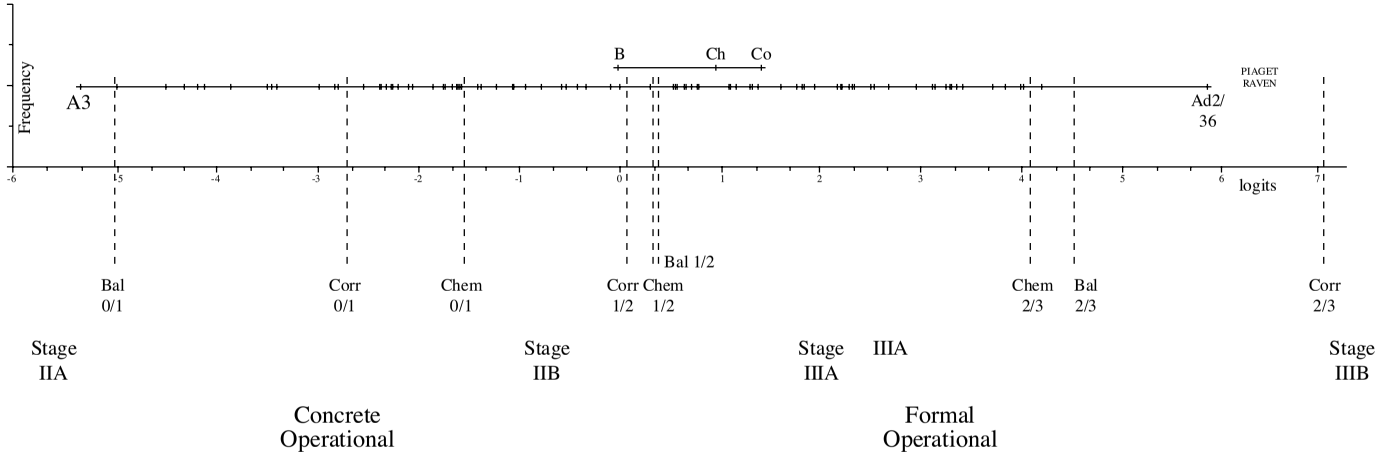
As can be seen, the Piagetian items are of different difficulties from one another and are ordered as expected with the proportionality (Bal) task being the easiest and the correlational task (Corr) being the most difficult. Relative to the spread of difficulties in the RPM, they are fairly similar to one another in average difficulty and correspond to the middle of the range locations of items on the RPM.

Because the transition points between stages of the cognitive developmental tasks are in the same metric as the locations of the items, it is possible to plot the category characteristic curves for the three Piagetian tasks on the same scale and then examine where the transition points occur relative to the RPM item locations (as well as the ability estimates). These curves, which can be superimposed on the scale of item locations, are shown in Figures 2.2a, 2.2b, & 2.2c for the Bal, Chem and Corr tasks, respectively. Note that the transition points between stages are at different locations on the continuum for the three tasks and that the transition points between stages IIA/IIB and between stages IIIA/IIIB are much more varied, relative to one another, than are those between stages IIB and IIIA which are at virtually the same location for each of the three tasks.

Discussion

This study demonstrates the feasibility and usefulness of integrating two basic approaches to the study of the development of intellectual functioning. Using Rasch's extended logistic model, it is possible to quantify variables from the two perspectives and to scale them jointly onto a continuum that has a consistent unit of measurement throughout the operating range of the combined variables, enabling a direct comparison of the two types of items (tasks) and person performances on them. This, in turn, helps illuminate our understanding of the variables being examined: in this

Figure 2.1. **Joint Scaling of RPM and Three Piagetian Tasks**



case, the results indicate the two variables are different manifestations of the same latent trait which is postulated to be that of abstract reasoning. This supports the notion of Snow, Kyllonen, and Marshalek (1984) of the centrality of the RPM as a measure of general intellectual functioning. It also supports a similar notion concerning the centrality of Piagetian tasks as measures of the same construct.

The fact that, in the RPM, the earlier perceptual items scale together with the later more "analytic" ones also indicates that there is

Figure 2.2a. **Category Characteristic Curves for Chemical Task**

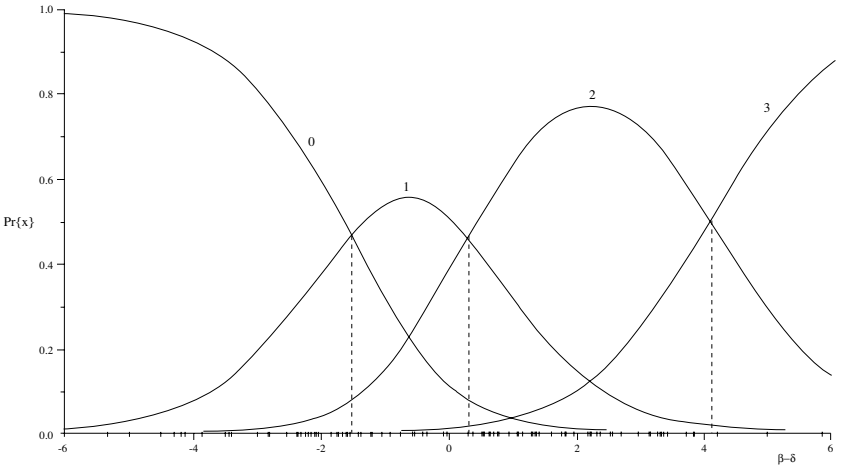


Figure 2.2b. **Category Characteristic Curves for Balance Task**

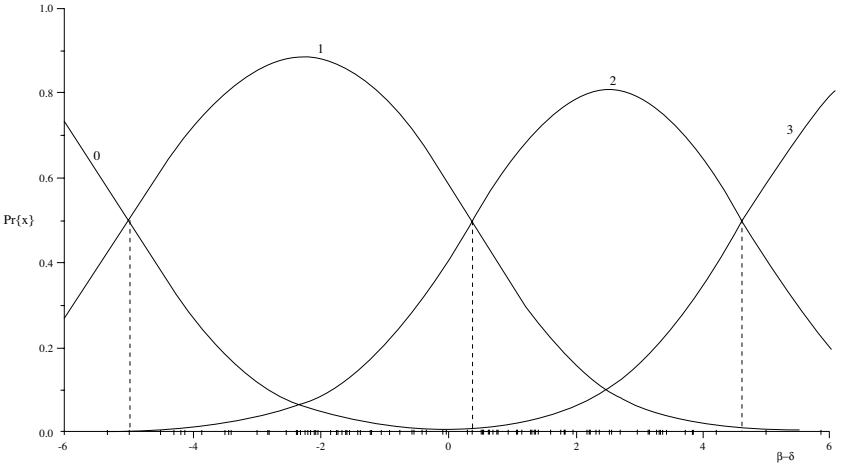
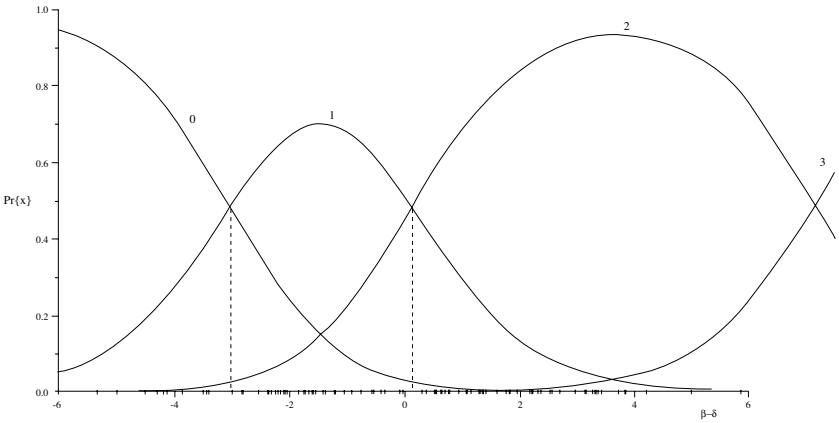


Figure2.2c. **Category Characteristic Curves for Correlation Task**



something common to both those types of processing, that is, the one type (perceptual processing) is a prerequisite for being able to use the other (analytic processing), and *they are both manifestations of a similar latent trait – that of abstract reasoning.* This conceptualisation fits well with the Piagetian sequence of stages in which preoperational thought (in which the person tends to interpret the world in terms of perceptual information) predates, and is a prerequisite for, operational thought in which the person is not at the mercy of purely perceptual information but is able to reason according to what the person *knows* to be the case rather than what s/he *sees* is the case.

Examining the relationships amongst the two operationalised variables of interest, it can be seen that they cover a wide – and a similar – range of ability, and that the top of the RPM coincides fairly well with the attainment of the second stage of formal reasoning (Stage IIIB) in two of the three tasks which were used. (The attainment of formal operations in the third task – Correlations – occurs beyond the location of the most difficult of the RPM items.)

The results show that these Piagetian tasks (and presumably others) are of different difficulties. The category characteristic curves indicate the probabilistic nature of stage categorisation and also show the possibility of stages for different tasks not being identical in meaning so that operating at a particular stage according to one task does not necessarily indicate the ability to do so at the same stage on another task for any one person. Considering this in more detail, it is evident from Figure 2.1 that although

the transition points for the three Piagetian tasks occur at different positions on the continuum, this is so only for the transition points between sub-stages – the transition points for all three tasks between stage II and stage III are very close together. This indicates, firstly, that people develop the reasoning processes required for different tasks at different times, and thus some tasks are more difficult at that stage than others are. For example, it is more difficult to pass from IIA to IIB in the Chem than in the Bal task – at one level of scale, the processes required for the Bal task at this stage would seem to be acquired earlier than those for the Chem task. This is consistent with Flavell's explanation of asynchrony in development on different tasks being due to tasks requiring different levels of functional maturity (Flavell, 1971). Secondly, the close proximity of the IIB/IIIA thresholds for all three tasks suggests that there is a more significant, major transition point between stages II and III: it seems people need to develop thinking processes that are common to all the tasks before they can change from stage II to stage III. This supports Piaget's theory (Piaget, 1972) that people may exhibit discontinuity in developing reasoning processes for different tasks: however, despite this, the transition between stages II and III (concrete to formal operations) remains a major qualitative change that is common to many tasks. This interpretation also supports the contention of Fischer, Pipp, and Bullock (in Kitchener, Lynch, Fischer, & Wood (1993) in addressing the development of reflective judgment, that when a new developmental level is emerging, people spurt in a whole range of domains. If this is so, then development would appear more stage-like at major transition points than between these points when different skills are developing at different times and at different rates (that is, people might be spurting in a specific area at different times). It would seem, then, that with regard to a specific task, development would appear more stage-like (in that minor spurts would be more obvious) than if development is considered at a more general level when spurts in different tasks coincide to a greater extent. This means that the degree to which stage-like development is obvious will depend on the level of scale at which measurements are made, and the relationship of the scale to the tasks one is using.

In regard to the expected ordering of the three tasks (Chem, Bal, Corr), it would appear that, although this order may vary through the development of concrete operations, for the firm establishment of formal operations (the transition from IIIA to IIIB), the processes required by Chem (full understanding of the combinatorial – Stage IIIB) need to be in place before those for Bal are fully established (negation and

reciprocity) and, similarly, those for Bal need to be in place before Corr is fully established. Thus, for the full attainment of formal operational thinking (stage IIIB), the order of establishment of reasoning processes seems to be in the order indicated by Inhelder and Piaget (1958), and Piaget's contention that the combinatorial underpins formal operational, propositional reasoning is supported.

Overall, the mapping of these two different variables – one from the quantitative, psychometric tradition and one from the qualitative, cognitive-developmental tradition, indicates that, at one level, quantitative and qualitative change are closely interlinked – the one *is* the other. Small quantitative increments eventually result in a major qualitative transformation of thought processes.

It is useful here to use an analogy to an evolutionary, biological process. Certain anatomical structures in early insect forms have the function of heat-regulation. However, above a critical body/wing ratio, these structures function, rather, as wings. Thus, more of the same results, not in a simple quantitative accumulation, but in a "complex reordering of parts with invention of new items" (Gould,1991). With respect to intellectual development, similarly, a sequence of simple and small quantitative changes translates into a major alteration of quality.

PART II

This Part of our paper seeks to extend the work summarized above by identifying parallels between psychometric and cognitive-developmental approaches and thus investigate the cognitive processing involved in solving the RPM items which correspond to the transition from concrete to formal operations on the Piagetian tasks.

A Taxonomy of Matrix Items

Central to the study is the characterisation of the algorithms or rules that a person needs to educe in order to solve RPM items. Previous work in the area includes that of Green & Kluever (1991), Hornke & Habon (1986) and Carpenter, Just, & Shell (1990). However, because these characterisations seemed to have disadvantages, this paper embodies an approach which differs from the others in several respects.

There have been several attempts at categorising matrix-like test items (particularly, the items of the different forms of the Raven Progressive

Matrices) either in terms of the elements of the items (the structure) (e.g. Green & Kluever, 1991, identified structural characteristics of the Coloured Progressive Matrices (CPM) such as symmetry/asymmetry, vertical/horizontal, straight/curved lines, number of dimensions), or the algorithms (the rules or cognitive operations) required to solve them (e.g. Hornke & Habon, 1986; Carpenter, Just, & Shell, 1990). Raven, himself (Burke, 1958), categorised the items of the SPM according to five basic characteristics related to the algorithms required to solve the items – one for each of the five sets A to E as shown in Table 2.1.

The existing taxonomies were perceived to have disadvantages: firstly, it seemed several levels of analysis of items were needed: the existing taxonomies were either too broad or too detailed to be useful for this study; secondly, it seemed useful to combine the structure of an item and the (theoretical) processes or operations required to solve it; and, thirdly, some of the nomenclature used in previous research has not been strictly correct, (for example, "identity", in regard to a matrix, is not a repetition of identical figures in each cell of the matrix). Hence, a new taxonomy was developed to incorporate several levels of analysis, first, the Matrix size (1x1, 2x2, or 3x3); second, the five general Algorithms and (theoretical) processes associated with items; third, the Operators by which the algorithms are broken down to a more specific levels which describe detailed rules applying to the horizontal and vertical axes of a matrix, and the direction in which they operate; and, finally, Attributes which take account of the structural elements of the items the types and number of shapes and/or patterns making up the elements of a matrix. The taxonomy is shown in Table 2.2.

Most of the levels shown in Table 2.2 have sub-levels within them. The levels of major importance for this paper are the Algorithms and their sublevels and the Operators which detail the workings of the algorithms with regard to process and direction (horizontal or vertical) and, therefore,

Table 2.1. Raven's Description of the Five Types of SPM Items

SPM Set	Type
A	Continuous patterns
B	Analogies between pairs of figures
C	Progressive alteration of patterns
D	Permutations of figures
E	Resolution of figures into constituent parts

these are described in more detail in Tables 2.3 and 2.4, respectively. The algorithm Continuous has subcategories indicating the intra-matrix structure which may be "Jigsaw" (symmetrical pattern in the horizontal and vertical), or which may be similar to one of the major algorithms such as Reflection or Seriation. The algorithm Distributions has subcategories indicating whether the elements (or subelements) are rotating through the matrix or not, as does the algorithm Transformations. The algorithm Equation has subcategories indicating whether the equation is a union or an intersection and, within each of these, whether the elements (or subelements) are segregated or integrated.

The classification for each of the items of the SPM and APM (Sets I and II) at one level only (Algorithm) is shown in Table 2.5. The classification of the items was carried out by the researcher and a trained assistant in order to establish the reliability of the classification system and agreement was 100%.

In theory, respondents would have to use the operations associated with the algorithms in order to solve the items. Respondents may not always do this, for instance, a more perceptually-based, less abstract process may be used, but this would be considered a lower-order process which would be successful only with the easier items in a group of items using a particular algorithm.

Table 2.2. Hierarchical Taxonomy for the Raven's Progressive Matrices

Category 1 MATRIX SIZE	Category 2 ALGORITHM
1 x 1	Continuous ("jigsaw") (C)
2 x 2	Reflection (R)
3 x 3	Seriation (S)
	Distribution (D)
	Transformation (T)
	Equation (union or intersection) (EU or E)

Category 3 OPERATORS (on attributes)	Category 4 ATTRIBUTES
Same/different(across horizontal and vertical axes)	Whole/part elements
Operators, e.g. increasing number or size, rotation of elements	Shape e.g. square, cross
Direction e.g. left to right, top to bottom	Pattern, e.g. solid black, vertical stripes
	Number of elements or subelements

Table 2.3. Sublevels of the Category Algorithms in the Taxonomy for RPM Items

Algorithm	Subalgorithm
1. Continuous	a) "Jigsaw" b) Reflection c) Seriation d) Rotational distribution e) Equation
2. Reflection	No subcategories (differences occur at the level of Attributes)
3. Seriation	No subcategories (differences occur at the level of Attributes)
4. Distribution	a) Different elements – nonrotational Different elements – rotational b) Transformation of elements – nonrotational Transformation of elements – rotational
5. Equation	a) Union – segregated Union – integrated b) Intersection – segregated Intersection – integrated

Piagetian Stages

Although the formulation of Piaget's stages are well-known, to make the connections between psychometric and Piagetian theory in a later section, the characteristics of the stages are summarised here.

Piaget postulated that the development of cognitive functioning occurs in stages characterised by different modes of thinking or reasoning about and interpreting the world. Stages are universal, invariant and age-related.

Sensori-motor stage: This mode of interpreting the world is not addressed in this paper.

Pre-operational stage (Stage I): The child tends to interpret the world through perceived appearances rather than through inferred reality (Flavell, 1977), that is, s/he interprets the world in terms of what objects/situations look or sound like rather than in terms of an underlying reality: for a preoperational child, perceptual data *is* reality. Piaget used the term "figurative" to refer to such activities – activities which "represent reality as it appears, without seeking to transform it" (Piaget, 1970, p717).

Concrete-operational stage (Stages IIA and IIB): In contrast to the above, at this stage, a child interprets the world in terms of an "underlying

Table 2.4. Sublevels of the Category Operators in the Taxonomy for RPM Items

Operators *		Description
1. Increasing		number, size, amount or position
2. Form change		e.g. Three different shapes or patterns in a 3x3 matrix
3. Reflection		reflection of patterns and symmetrical/ asymmetrical figures
4. Rotation		rotation of element positions across rows/ columns of matrix
5. Transformation		sequential transformation of structure/quality of element the "2nd Col" rule: the 2nd column element gives the rule for transforming the 1st column
6. Union:	Segregated	overlap 1st and 2nd columns: retain all subelements
	Integrated	
7. Intersection:	Segregated	a) overlapping elements cancel if similar $(A \leftrightarrow B)' = C$
	Integrated	b) nonoverlapping elements remain $(A \leftrightarrow B)'' = C$
		c) overlapping elements remain $(A \leftrightarrow B) = C$
		d) one element cancels others even when it is different
* Operators 1 and 4 are also coded according to the Direction of operation as follows:		
Horizontal plane:		Left to Right or Right to Left Top to Bottom or Bottom to Top
Vertical plane:		Clockwise or Anticlockwise

reality", or rule-governed ways of thinking (Howes, 1990). The child develops a logic of classes, relations and number, that is s/he reasons in terms of these objects, but cannot link any one object to any other except those that are adjacent to it (Piaget, 1972). The concrete-operational child can reverse actions, that is, realise that one action can reverse or nullify the opposite one, but is not yet able to link the two types of reversibility (negation and reciprocity). The child is limited by what is empirically given, that is, s/he is tied to reasoning about one particular situation at a time (hence the term "concrete" thinking).

Formal operational stage (Stages IIIA and IIIB): Adolescents at this stage differ from concrete thinkers in several ways. Firstly, they can

Table 2.5. All RPM Items in Increasing Order of Difficulty and Their Classification by the Taxonomic Category 'Algorithm'

Number	Difficulty (in logits)	Name	Algorithm	Number	Difficulty (in logits)	Name	Algorithm
1	-5.35	A3	C (J)	53	0.30	AdII/4	S
2	-5.01	A5	C (J)	54	0.30	E2	E
3	-4,52	B3	R	55	0.38	D10	D
4	-4.32	B2	R	56	0.52	AdII/11	E
5	-4.20	A6	C	57	0.54	E5	E
6	-4.13	A7	C	58	0.56	AdII/12	E
7	-3.86	A9	C	59	0.63	AdII/10	S
8	-3.50	D2	D	60	0.64	AdI/12	E
9	-3.45	AdI/4	C	61	0.65	B12	R
10	-3.40	B4	R	62	0.70	E4	E
11	-2.98	B5	R	63	0.76	AdII/14	S
12	-2.83	B6	R	64	0.78	C11	S
13	-2.79	AdI/1	C	65	1.08	AdII/16	E
14	-2.54	D5	D	66	1.09	AdII/17	T
15	-2.38	C7	S	67	1.15	AdII/13	S/D
16	-2.36	A10	C	68	1.29	AdII/15	T/E
17	-2.31	C3	S	69	1.32	AdI/9	S/T
18	-2.24	C2	S	70	1.36	AdII/8	D
19	-2.24	C5	S	71	1.37	E6	E
20	-2.24	B10	R	72	1.60	C8	S
21	-2.20	AdI/2	C	73	1.76	E7	E
22	-2.16	D3	D	74	1.81	AdII/19	E
23	-2.10	B11	R	75	1.84	AdI/11	T/E
24	-2.09	B9	R	76	1.94	AdII/20	E
25	-2.08	C9	S	77	2.17	AdII/22	E
26	-2.05	AdI/3	C	78	2.21	AdII/21	S/D/T
27	-1.86	B7	R	79	2.21	E8	E
28	-1.76	A8	C	80	2.29	D11	D/T
29	-1.71	B8	R	81	2.32	AdII/18	T
30	-1.73	D4	D	82	2.35	AdII/23	E
31	-1.67	C4	S	83	2.51	E10	E
32	-1.62	A11	C	84	2.55	D12	S/D/T
33	-1.61	D8	D	85	2.69	E9	E
34	-1.60	A12	C	86	2.96	AdII/26	D/S/T
35	-1.57	AdI/7	D	87	3.12	AdII/33	E
36	-1.42	AdI/5	S	88	3.15	AdII/30	T
37	-1.38	D6	D	89	3.26	AdII/31	S
38	-1.23	AdI/6	S	90	3.30	AdII/24	S
39	-1.22	C6	S	91	3.32	AdII/27	D/T

Table 2.5. All RPM Items in Increasing Order of Difficulty and Their Classification by the Taxonomic Category 'Algorithm' *(continued)*

Number	Difficulty (in logits)	Name	Algorithm	Number	Difficulty (in logits)	Name	Algorithm
40	-1.06	AdI/10	E	92	3.36	AdII/25	S
41	-1.05	D7	D	93	3.43	AdII/34	S/D/T
42	-1.04	AdII/5	S	94	3.71	C12	S
43	-0.93	AdII/6	S	95	3.83	E11	E
44	-0.77	AdI/8	S/D	96	3.84	AdII/28	S/D/T
45	-0.57	AdII/9	E	97	3.98	AdII/29	T
46	-0.53	AdII/1	D	98	4.01	AdII/35	E
47	-0.42	E3	E	99	4.01	E12	E
48	-0.41	D9	D	100	4.12	AdII/32	S/T
49	-0.41	AdII/3	S	101	5.86	AdII/36	E
50	-0.33	AdII/2	S				
51	-0.08	AdII/7	E				
52	0.01	C10	S				

conceive of not only the real and actual, but the possible, that is, they can go beyond reality to consider all theoretically possible situations associated with a task. Secondly, they tend to solve problems using hypothetico-deductive reasoning (Piaget, 1972). Thirdly, in contrast to children's intra-propositional thinking, the formal operator can reason inter-propositionally: s/he can consider the logical relationships between propositions (relations amongst relations) rather than the factual relationship between a proposition and reality only. This kind of logic is based on the use of the combinatorial system (the sixteen binary propositions of logic) which is the basis of combinatorial and permutational analysis, and the INCR group (Identity, Negation, Reciprocity and Correlation) which requires the combination of operations (Piaget, 1972; Bond, 1980). This constitutes an ability to reason at a second, or higher order level. Piaget saw the development of these abilities as relatively continuous from concrete to formal operations, but thought that development would be particularly rapid from about 12 years onward (Inhelder & Piaget, 1958).

All abilities available during each stage become consolidated during subsequent developmental stages, that is, concepts become more stable and robust as development proceeds.

Results and Discussion (Part II)

The joint scaling of the combined Matrices items and Piagetian tasks has already been shown graphically in Figure 2.1. The transition points and their significance have also been addressed in Part 1, as has the ordering of the Piagetian tasks. We now discuss, therefore, the ordering of the RPM items according to the taxomony provided; the relationships between the processing used in these items and the processing characteristics of the Piagetian stages; and, finally, the relationships between the RPM items and the transition points between Piagetian stages.

Ordering of the RPM Items

With respect to *the taxonomy of items*, although items from different taxonomic categories exhibit a range of difficulties, thereby resulting in the categories overlapping in location on the continuum, in general, there is a distinct sequence in the occurrence of certain algorithms at particular locations on the continuum in the order of difficulty as expected from Raven's original conception of the order and the taxonomy used here, that is, Continuous ("jigsaw" and other), Reflection, Seriation, Distribution (with and without rotations and/or Transformations), Equations (unions), and Equations (intersections). From a study of the order of the items both within and between these algorithms, it would be possible to identify aspects of processing which make an item more or less difficult to solve. The paper does not address this issue systematically and specifically, however, one aspect that is particularly relevant here is the number of algorithms involved in solving the items: the easiest items employ only one algorithm, and items employing two and then three algorithms are increasingly difficult. But there are very difficult items which employ only one algorithm, thus difficulty of items is related to either the type of algorithm or the number of algorithms, or both the type and the number.

Order of RPM Algorithms and Piagetian Tasks

The ordering of the algorithms for the items clustered around the transition points for all tasks is consistent in that the easier algorithms are associated with the earlier transition points and the algorithms become progressively more difficult at each succeeding transition point. This, again, supports the convergence of the two approaches to intellectual

functioning: at a general level, one reflects the other. The ability to cope with items that employ two algorithms increases gradually through the stage of concrete operations, but the ability to deal with items using more than two algorithms is available only once formal operations are firmly established – almost at the transition between the substages IIIA and IIIB.

Based on these results, the answer as to whether the successful solution of matrix problems requires the attainment of formal operations or simply concrete operations, is, therefore, that it depends what the matrix problems are: concrete operational thinking are sufficient for the solution of matrix problems where the algorithm is fairly simple (one of two algorithms – Continuous or Reflection), or a easy example of a more difficult algorithm (Seriation or non-rotational Distribution with no transformations) which involve not more than two algorithms, both of which have to be relatively simple. During the stage of concrete operations, the ability to use operations simultaneously is quite limited. This supports Hubbs-Tait's conclusion that matrix items involving the interaction between variables require formal operational thinking (Hubbs-Tait, 1986).

Characteristics of RPM Items at Transition Points

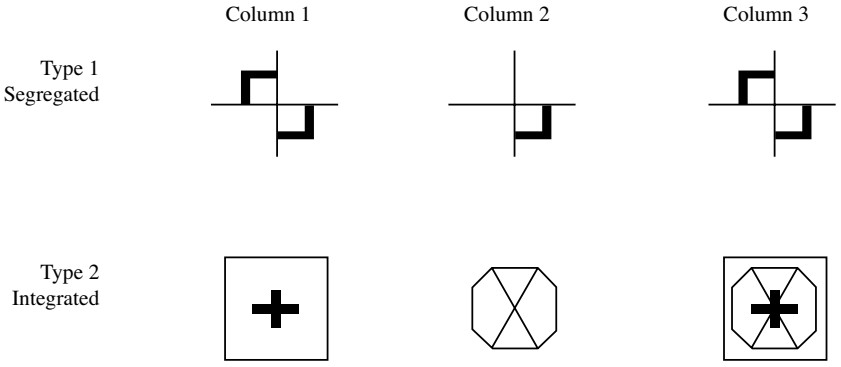
Table 2.6 shows the RPM items that occur within two standard errors above and below each of the thresholds between stages and sub-stages for each of the Piagetian tasks and the type of algorithm associated with each item according to the taxonomy presented in Table 2.3. It is evident that items occurring in the vicinity of each of the thresholds are characterised not only by particular algorithms but, further, usually include the first instances (in order of difficulty) of these algorithms. The four transition points are now addressed in turn.

First, considering the RPM items between the two major stages, II and III, the outstanding group consists of items which use the E algorithm (unions and intersections), in particular, for the first time, those items described as integrated unions where elements are superimposed directly on one another by joining the first and second columns (or rows) to form the third column (or row). No instances of E (intersection) occur before this transition. Three instances of the E (union) items occur before this point in the scale, they are, however, the first of the segregated unions: much simpler, than the integrated unions at a finer level of analysis in that their construction is such that joining the first two elements of the matrix results in a third element in which the subelements of the first two elements are still distinct from one another rather than overlapping one another completely to form a figure that looks very different from

Table 2.6. RPM Items (in Order of Difficulty) Occurring at the Thresholds Between Piagetian Stages, and Their Taxonomic Classification by Algorithm

Stage	Piagetian task	Location (in logits) +2 std errors	Continuous	Reflection	Algorithm Seriation	Classification Distribution	Transformation	Equation U	Equation I
I	Chem								
	Bal								
	Corr								
IIA/ IIB	Bal	-5.01 + 1.00	A3,A5,A6,A7	B3, B2					
	Corr	-3.03 + 0.76	Ad1/4,Ad1/1, A10	B4, B5, B6	C7, C3	D2, D5			
	Chem	-1.53 + 0.50	A8, A11, A12	B7, B8	C4, Ad1/5, Ad1/6, C6	D4, D8, Ad1/7, D6			
IIB/ IIIA	Corr	0.12 + 0.34			C10, Ad2/4	D10		Ad2/7, E2	
	Chem	0.30 + 0.34			C10, Ad2/4 Ad2/10	D10		E2, Ad2/11	E5, Ad2/12
	Bal	0.35 + 0.34		B12	Ad2/4, Ad2/10	D10		E2, Ad2/11	E5, Ad2/12, Ad1/12
IIIA/ IIIB	Chem	4.07 + 0.40			C12		Ad2/28		E11,Ad2/35, E12, Ad2/32
	Bal	4.58 + 0.52					Ad2/29		AD2/35, E12AD2/32
	Corr	7.14							

Figure 2.3. **Examples of One Horizontal Line of the Matrices of the Two Types of the Algorithm Union – Segregated and Integrated**



its parts. (In fact, the subelements are reflections of one another and the ability to operate with reflections is one that occurs earlier at the beginning of stage II, concrete operations.) Examples of the differences between Union (segregated) and Union (integrated) figures are shown in Figure 2.3.

The main difference between the Union (integrated) and Intersection (integrated) items and all earlier ones is that the former are the first items that require the respondent not only to go beyond the material presented in that they constitute *equations* involving the elements of the matrix, but also require a respondent to *know this explicitly*.

On the other hand, with items of the Union (segregated) type, it is possible to arrive at the correct answer simply because it "looks right", that is, one can rely on what Piaget described as a "figurative" way of thinking (producing what is given rather than having to transform the information in some way – the "logico-mathematical" way of thinking).

The joining (or separation) of the first two elements of the stem matrix in order to obtain the third element seems to be a second-order process which is not inherent in the perceptual properties of the stem matrix as is, for example, the redistribution of components across the elements of the stem matrix. This interpretation is consistent with Piaget's description of the characteristics of formal thinking as including an ability to deal with second-order relationships and more abstract ways of thinking and not being tied by what is immediately known or expected from the given perceptual task characteristics.

Second, *the transition from IIA to IIB for Chem and Corr* involves, in addition to some of the more difficult Continuous and Reflection items,

the solution of Seriation (S) and Distribution (D) items for the first time. The IIA/IIB transition also involves items which employ two algorithms for the first time. Seriation and Distribution items involve either seriation or some of the possible combinations or permutations of elements making up the matrix. Further, combinations are classifications of all possible classifications and permutations are seriations of all possible seriations – they are operations on operations (Piaget, 1970). Thus, the fact that the IIA/IIB transition items for Chem and Corr involve the Seriation (S) and Distribution (D) algorithms for the first time fits well with the expected ability to begin to use elementary combinations which is a hallmark of these Piagetian tasks at this level.

The ability to solve D items is consolidated in the formal operational stage, but that, at the same time, the ability to deal with transformations (in association with rotations) develops. Transformations items are of the type where one figure is transformed in some way across the elements of the matrix (rather than the elements being comprised of three distinctly different shapes).

Note, too, that the Seriation algorithm is consolidated across concrete to formal operations, but that the major novel Seriation item type that occurs for the first time at the formal operational stage is one in which *position* (rather than number, size or amount) is the characteristic of the subelement that is seried, so that subelements appear to be moving across (or down) the matrix.

Third, the same *transition point (Stage IIA to Stage IIB) for the Bal task* is associated with a focus on R items (reflection): there is a clear similarity here between the operations involved in reflection and the ability to coordinate what is happening on both arms of the balance in the Balance task (negation and reciprocity). This is elaborated during concrete operational stage and is complete before the advent of formal operations. An interesting exception to this is the occurrence of B12 at the formal operations level Bal task). The interpretation here is that B12 has been misclassified – although it uses reflection, it is possible that it is a forerunner of an Intersection algorithm item in that there are two subelements, each of which is present or absent alternately in either the horizontal or the vertical. It is hypothesised that some of the processing abilities required for dealing with intersections are needed to be successful with this item.

Fourth, and lastly, the *transition from IIIA to IIIB is considered for Chem and Bal* only since no items occur in the vicinity of the corresponding Corr threshold which occurs well beyond the most difficult item of the

RPM. The main group of items here is comprised of intersection items where only non-overlapping elements between columns 1 and 2 remain in column 3 (or row 3). In particular, Intersection where two *different* patterns (or shapes) may eliminate each other, rather than the integration of two similar shapes resulting in their elimination . Again, this seems to fit well with the extension to being able to consider all possibilities and to coordinate several different operations simultaneously, which are hallmarks of the formal operational stage.

Another group of items occurring at this transition point is Transformations where the respondent has to recognise more than one type of transformation *and use them simultaneously*. In addition, other algorithms are involved. For these items, it is not possible for a respondent to regard transformations on elements simply as totally different figures (as it is in earlier Transformation items) – the types of transformation have to be recognised as such and coordinated with each other together with the use of multiple algorithms.

The ability to recognise the use of the same figure in different forms, to see the dependence amongst them even when more than one transformation is used within one figure, and take account of other operations at the same time, fits well with the notion of an intellect whose horizons are expanding to include all possibilities and which can deal with many possibilities at one time (again, an example of being able to perform operations on operations).

Another characteristic is that items which employ three algorithms start to occur just before this transition point.

No items occur in the vicinity of this transition point for Corr, however, it would be possible to predict what kind of items might do so, given their locations and what is known about the processes required at these levels in the Piagetian tasks.

Conclusions to Part II

The results of this study indicate the close relationship between Piagetian changes in modes of thinking and the processes associated with the solution of different types of psychometric items.

They also support the Piagetian notion of a major change in quality of thinking from concrete to formal operations and indicate that this major change is reflected, too, in the ability to solve psychometric items

requiring particular types of processing. Changes involve the ability to recognise multiple aspects of a problem and to coordinate them, and to function less in a figurative (literal) and more in a logico-mathematical (abstract) way. This, in turn, strengthens the view that quantitative and qualitative development are intimately related to one another rather than being distinctly different from one another: they may be considered as being at different levels of scale, that is, the psychometric conception is simply a finer level of scale than the Piagetian one. It does not follow, however, that quantitative measures always at a finer level than qualitative – both may be found at all levels of scale: they are inseparable.

A further question that can be investigated through the use of this data is the relationship between the Piagetian transitions and intellectual growth spurts as measured on a psychometric variable. A major growth spurt, common to both sexes, has been demonstrated to occur during puberty, beginning for some children at the age of 11 years, and being completed by virtually all children by age 15 years (Andrich & Styles, 1994). This is entirely in accord with Piaget's postulation of a major qualitative change in reasoning from concrete to formal operations within a similar age range.

Overall, the conclusion is that the integration of the psychometric and cognitive-developmental approaches to the study of intelligence allows a deeper understanding of the variables under investigation and their relationships to one another.

References

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon-Tuma (Ed.), *Sociological methodology.* San Francisco, CA: Jossey-Bass.

Andrich, D., & Constable, E. (1984). Studying unfolding developmental stage data based on Piagetian tasks with a formal probabilistic model. Symposium on studies of Piagetian cognitive stages. *Annual Meeting of the American Educational Research Association, New Orleans.*

Andrich, D., Lyne, A., & Sheridan, B. (1990). *ASCORE: a manual of procedures.* Western Australia: Murdoch University.

Andrich, D., & Styles, I. (1994). Psychometric evidence of intellectual growth spurts in early adolescence. *Journal of Early Adolescence, 14(3)*, 328-344.

Bond, T. (1980). The psychological link across formal operations. *Science Education, 64(1)*, 113-117.

Bond, T. (1993). Empirical research and Piagetian theory: Quantitative approaches applied to qualitative theory, *Paper presented at the Annual Australian Association for Research in Education, November, Fremantle, Western Australia.*

Brainerd, C. J. (1978). The stage question in cognitive-developmental theory. *Behavioral and Brain Sciences, 2*, 173-213.

Burke, H. R. (1958). Raven's Progressive Matrices: A review and critical evaluation. *Journal of Genetic Psychology, 93*, 199-228.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97,(3)*, 404-431.

Carroll, J. B., Kohlberg, L., & De Vries, R. (1984). Psychometric and Piagetian intelligences: toward resolution of controversy. *Intelligence, 8,* 67-91.

Case, R. (1978). Intellectual development from birth to adulthood: A neo-Piagetian interpretation . In R. S. Siegler (Ed.), *Children's thinking*: *What develops?* (pp. 37-72). Hillsdale, NJ: Erlbaum.

Chomsky, N. (1976). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Choppin, B (1968). An item bank using sample-free calibration. *Nature, 219,* 870-872.

Choppin, B. (1983). *A fully conditional estimation procedure for Rasch model parameters. Report No.196*. Centre for the Study of Evaluation, Graduate School of Education, University of California, Los Angeles.

Davison, M., King, P. M., Kitchener, K. G., & Parker, C. A. (1980). The stage sequence concept in cognitive and social development. *Developmental Psychology, 16,* 121-131.

De Vries, R. (1974) Relationships among Piagetian, IQ and achievement assessments. *Child Development, 45*, 746-756.

De Vries, R., & Kohlberg, L. (1977). Relations between Piagetian and psychometric assessments of intelligence. In L. Katz (Ed.), *Current topics in early childhood education.* (Vol. 1). Norwood, NJ: Ablex.

Duncan, O. D. (1985). Probability, disposition and the inconsistency of attitudes and behaviour. *Synthese*, *42*, 21-34.

Elkind, D. (1974). *Children and adolescents*: *Interpretive essays on Jean Piaget (*2nd Edition). London: Oxford University Press.

Farnham-Diggory, S. (1972). *Cognitive processes in education*: *a psychological preparation for teaching and curriculum development.* New York: Harper and Row.

Fischer, K. W., Pipp, S. L., & Bullock, D. (1984). Detecting discontinuities in development: method and measurement. In R. Harmon and R. Emde (Eds.), *Continuities and Discontinuties in Development* (pp. 95-122). New York: Plenum Press.

Flavell, J. H. (1971) Stage-related properties of cognitive development. *Journal of Cognitive Psychology, 2,* 421-453.

Flavell, J. H. (1977). *Cognitive development.* New Jersey: Prentice-Hall.

Gould, S. J. (1991). *Bully for Bronasaurus*: *Reflections in natural history.* London: Random Century Ltd.

Green, K. E., & Kluever, R. C. (1991). Component identification and item difficulty of Raven's Progressive Matrices items. *Paper presented at the annual meeting of the National council of Measurement in Education, Chicago, April.*

Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10(4)*, 369-380.

Howes, M. B. (1990). *The psychology of human cognition*: *Mainstream and Genevan traditions.* New York: Pergamon Press.

Hubbs-Tait, L. (1986). Transitions in the reasoning of pre- and early adolescents: A new method of assessment. *Paper presented at the sixteenth annual symposium of the Jean Piaget Society, Philadelphia, May*.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* London: Routledge and Kegan Paul.

Kitchener, K., Lynch, C., Fischer, K. W., & Wood, P. K. (1994). Developmental range of reflective judgment: the effect of contextual support and practice on developmental stage, *Developmental Psychology*, *29,* 893-906.

Lim, T. K. (1988). Relationships between standardised psychometric and Piagetian measures of intelligence at the formal operations level. *Intelligence, 12*, 167-182.

MacKay, C. I., Fraser, J., & Ross, I. (1970). Matrices, three by three: classification and seriation. *Child Development, 41*, 787-797.

Philp, H., & Kelly, M. (1974). Product and process. *British Journal of Educational Psychology, 44*, 248 – 265.

Piaget, J. (1970). Piaget's Theory. In P. Mussen, *Carmichael's handbook of child psychology.* New York: Wiley.

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human development, 15,* 1-12.

Piaget, J., & Inhelder, I. (1958). *The growth of logical thinking from childhood and adolescence.* London: Routledge & Paul Ltd.

Rasch, G. (1960/80). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research), expanded edition (1980). with foreword and afterword by B. D. Wright. Chicago, University of Chicago Press.

Raven, J. (1989). The *Raven Progressive Matrices*: A review of national norming studies and ethnic and socioeconomic variation within the United States. *Journal of Educational Measurement, 26(1),* 1-16.

Raven, J. C. (1940). Matrix tests. *Mental Health, 1*, 10-18.

Seigler, T. J., & Richards, L. (1991) The NUD.IST qualitative data analysis system. *Qualitative Sociology, 14*, 307-324.

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence,* Vol. 2 (pp. 47-103). Hillsdale NJ: Erlbaum.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Styles, I. (1991). Clinical assessment and computer-adaptive testing, *International Journal of Man-Machine Studies, 35,* 133-150.

Styles, I., & Andrich, D. (1994). Linking the standard and advanced forms of the Raven's Progressive Matrices in both the pencil-and-paper and computer-adaptive-testing formats. *Educational and Psychological Measurement, 53(4)*, 905-925.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, *16*, 433-51.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

PART II

Practical Measurement Issues: Lessons From 75 Years' work With *Item Response Theory*: Benefits, Problems, and Potential Solutions

In developing his *Progressive Matrices* (RPM) tests, J. C. Raven anticipated the development of *Item Response Theory* (IRT) in that he plotted what (technical disputes over terminology excepted) have since become known as *Item Characteristic Curves* (ICCS). These showed how the proportion of respondents getting each item right varied with total score. If the curves were irregular he tried to find the cause and correct (or, if necessary, reject) the item. He incorporated the curves for all the items into a single graph, so that he could see how closely the shapes of the curves corresponded to each other, whether they crossed over (implying that the order of difficulty varied with the ability of the respondents), and whether they were, as far as possible, equally spaced.

Although the logic for what he was trying to do was briefly explained in the test Manual (then known as the "Guide to the Use of" one or other of the tests) and elsewhere, the measurement model was not sufficiently differentiated from Classical Test Theory for most readers to appreciate just how distinctive it really was. This has only become clear to a significant number of people as a result of recent developments in Item Response Theory (IRT). Yet, although these developments have resulted in the logic of the approach being more widely understood, the fact that

the construction of the RPM was based upon them still generally passes unnoticed. Failure to appreciate just how different the measurement model deployed in the construction of the RPM was from classical test theory unfortunately resulted in some fairly widespread criticism of the tests stemming from attempts to apply procedures associated with classical test theory to evaluate the internal consistency of the RPM and to endless erroneous conclusions being drawn from research.

Only recently, by, with considerable difficulty, replicating Raven's methods using modern computer programs has it become possible to appreciate how close Raven had come to placing the scientific status of "eductive" ability, and the RPM as a measure of it, beyond dispute.

The chapters in this Section belatedly rectify these oversights.

The chapter by Anca Dobrean (née Domuta) describes the sampling procedures employed in the Romanian standardisation of the SPM ***Plus*** that yielded the data base on which most of the later chapters in this Section are based.

Raven, Prieler, and Benesch compare computer-generated Raven-type "empirical" ICCs with those produced using modern IRT programs. It emerges that the most widely applied version of IRT – the 1 parameter model – can yield results which seriously mislead researchers. Serendipitously, the research ends up demonstrating that both eductive and reproductive abilities are every bit as "real" as – and measurable in the same way as – high jumping ability or life expectancy.

The author's chapter summarising research conducted whilst a Romanian version of the *Mill Hill Vocabulary Scale* was being developed again reveals – perhaps in an even more striking way – that widely promoted IRT programs do not deliver the expected benefits. On the other hand, *Distractor* Characteristic Curves – i.e. plots of how the choice of each *wrong* answer varies with total score – yield information which is very useful to test developers. Beyond that, the chapter illustrates just how difficult it is to create a genuinely parallel version of what must be almost the archetypical form of IRT test – a vocabulary test made up of words of increasing difficulty.

The chapter by Prieler and Raven discusses the enormous methodological problems which arise in the thousands of studies which claim to measure and compare change – whether in groups or individuals – using more or less any test developed on the basis of Classical Test Theory … or even IRT-based tests which do not yield linear Test Characteristic Curves. Such test may, collectively, be described as being

grounded in "arbitrary *metrics*". But equally, if not more, serious errors in evaluation studies said to provide the basis for "evidence based treatment" (for example, in psychotherapy or education) stem from the adoption of what are best described as "arbitrary *measures*" … ie evaluation studies in which the researchers have concentrated their attention on only one or two outcomes (perhaps measured with highly reliable tests) instead of trying to get a rough fix on all potentially important outcomes – ie on the *comprehensiveness* of the evaluation. Both deficits can be overcome by adopting IRT-based procedures developed by Fischer and outlined in this chapter.

Chapter 3

The Need for, and Development of, the SPM *Plus*

John Raven

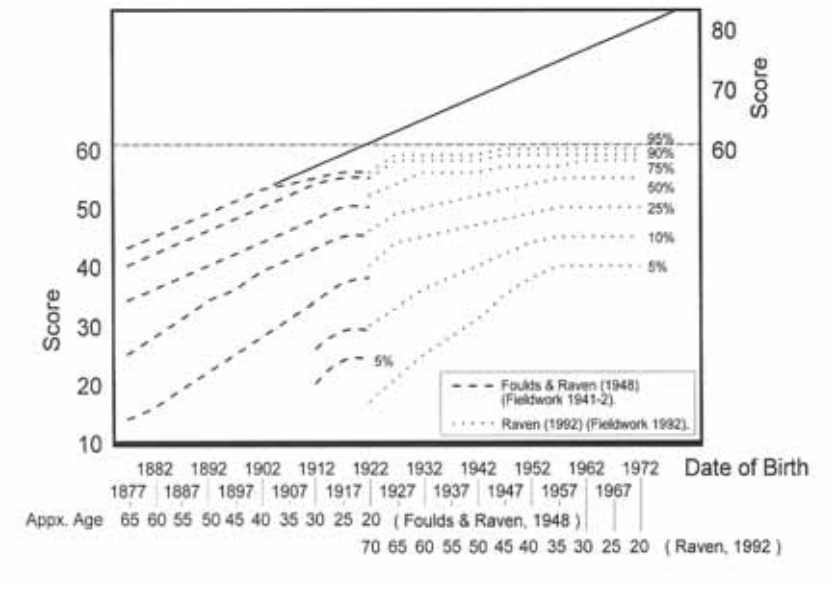
As we have seen, the development of the *Standard Progressive Matrices* ***Plus*** (SPM+) was precipitated by the dramatic and unexpected international increase in RPM scores that had taken place over the years. This resulted in the failure of the *Classic* Standard Progressive Matrices to discriminate above the 75th percentile among adolescents and young adults living in societies with a tradition of literacy.

The development of the SPM+ was, however, linked to the development of *parallel* versions of the both the *Coloured* and *Standard Progressive Matrices* tests – i.e. to the development of new tests in which the items would match those in the *Classic* versions on an item-by-item basis, both in overt solution strategy and in empirical difficulty. Only such tests would enable users to continue to refer to existing normative data with confidence and ensure that any new data they collected could form part of the international data pool which has proved so invaluable in documenting changes in test scores over time and between cultures.

Figure 3.1 plots the increase in SPM scores for adults born in each year from 1877 to 1972 and extrapolates the almost linear increase in the 95th percentile from 1877 to the point at which it begins to plateau (i.e. among those born in 1902) to a birth date of 1980. It shows that it would be necessary to introduce additional difficult items, and probably an 84-item test, to achieve the same discriminative power among those of higher ability born in 1980 as the *Classic* version had among those born before 1900.

Even a test of this length would not offer as much scope for increases above the 95th percentile as had (fortunately) been provided for in the *Classic* version. Consequently a test with about 90 items would be required to restore the discriminative power that the *Classic* SPM had among more able respondents in 1938.

Figure 3.1. *Standard Progressive Matrices*
100 years of Eductive Ability with Extrapolation of the 95th Percentile to 2000

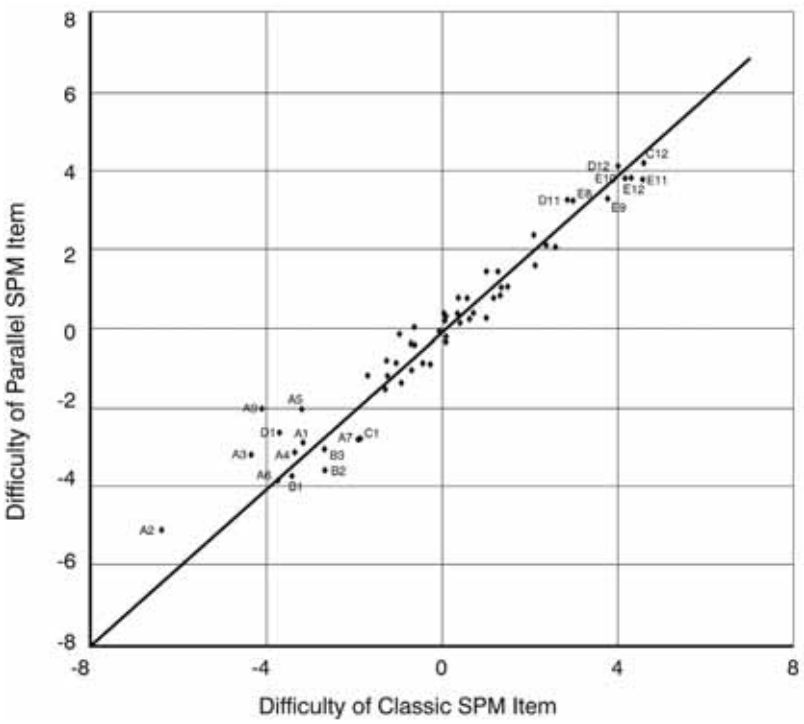


As described in Appendix 2 to the 1998-2004 editions of the SPM Section of the *Manual*[3.1], the energies of numerous people in several countries were harnessed to the task of developing the required items, conducting and analysing pilot studies, and finally testing the large number of people at all ability levels that were needed for an item-equating study.

Figure 3.2 plots the difficulty levels, expressed in logits, of the 60, new, parallel items against those in the *Classic* version of the SPM. It is clear that, with the possible exception of item A9, the difficulty levels of the items constituting the *Parallel* SPM closely match those they replace. Inspection of the parallel A9 revealed the reasons for the mismatch and the item was subsequently modified.

Turning now to the extension of the test to form the SPM+, 88 items were finally selected from a series of international pilot studies for inclusion in a very large international item-equating study, the design of which will be discussed in an Appendix to this chapter. Figure 3.3 shows the item difficulties of the 84 parallel and new items which remained after elimination of the four which had the poorest fit to a 1-parameter

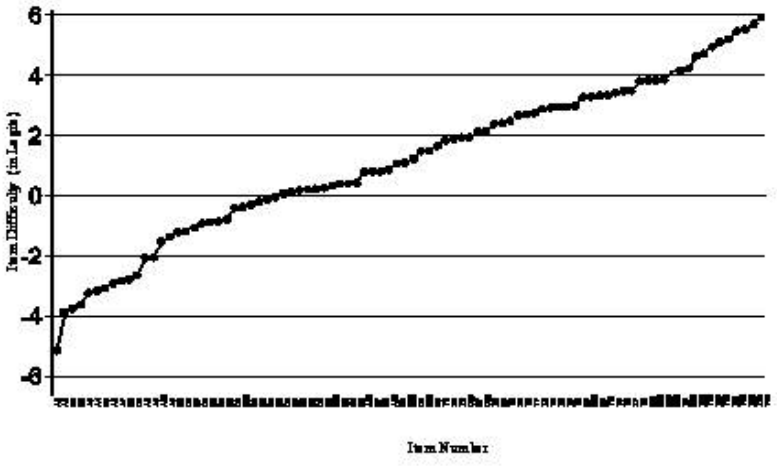
Figure 3.2. **Comparative Difficulties of *Classic* and *Parallel* Standard Progressive Matrices Items**
(Based on 1996 Item-Equating Study)



Item-Response-Theory model (this being most commonly referred to as the "Rasch model").

Although it is not immediately obvious from the graph once it has been reduced to a size suitable for inclusion here, inspection of a more detailed print out revealed that, in several sectors, there are a number of items having similar difficulty. It followed that, by eliminating alternate items in these areas, an almost linear increase in the difficulty of the items could be achieved. One of the sectors where the graph almost plateaus comprises items D3 to A11. Clearly, by eliminating 24 items, largely from those paralleling items from the original test, it would be possible to recreate a test having optimal length (in terms of fatigue and boredom) and yet discriminating across the entire range of intellectual ability. In fact, such a test would be a great boon since Carver[3.2] has shown that the use of tests in which total score does not increase directly with the

Figure 3.3. *Standard Progressive Matrices Plus*
1996 Item-Equating Study
Item Difficulties (in Logits) of Best 84 Items
(60 Parallel Items and 24 Additional Items)
arranged in order of difficulty.



difficulty of the most difficult item people are able to solve has led to serious misinterpretations of research findings. One example concerns apparent changes in the rate of maturation and decline of eductive ability with age. It is clear from Figure 3.3 that the distribution of items by difficulty is uneven. The result is that, when people work through the items contributing to plateau like that already mentioned, large increases (or decreases) in total score occur without commensurate increases or decreases in ability. This in turn results in rapid increases and decreases in raw score at certain ages that are not accompanied by accelerations or decelerations in actual ability. Yet the sudden increases or plateaux in raw scores at certain ages/ability levels has previously been interpreted to support the conclusion that there are leaps and plateaux in mental development when they are, at least in part, a measurement artefact.

Unfortunately, eliminating items to leave only those that result in equal increments in difficulty poses problems because each of the Sets in the *Classic* and *Parallel* versions of the SPM (i.e. A, B, C, D, and E) is made up of items of a different type. These not only require different forms of reasoning but also introduce those being tested to the logic required to solve the next most difficult item in that Set. Elimination of

the clearest candidates for removal would have resulted in a selection of 60 items which would have destroyed this unique property of the test. It would also have destroyed the comparability between the SPM and CPM. And it would have reduced the test's new-found ability to discriminate well among older adults and young children in zones where the 1938 version of the test did not work too well.

As a compromise, the items making up Sets A and B in the parallel test were left intact. For the *new* Set C, five items were selected (on the basis of both item difficulty and an examination of their logic) to represent the logical stages of each of the old Sets C and D and supplemented by two new items.

The difficulty levels of the items which remained are shown in a continuous graph in Figure 3.4 and, broken down by Set, in Figure 3.5.

It is apparent from Figure 3.4 that a reasonable approximation to a test made up of items having a linear increase in difficulty (assessed in logits) – and thus equal increases in total score for equal increases in ability – has been achieved without destroying the test's previously mentioned compatibility with the CPM and ability to discriminate among those with lower scores.

In summary, then, it would seem that, in developing the SPM+ we have achieved our objective of developing a test which restores the discriminative power at the upper end which the *Classic* SPM had when it was first developed and done this via a test which, like the *Classic* version. not only avoids boredom and fatigue, but also has more or less equal increments in item difficulty (once they have been arranged in difficulty order – which is not, however, the best order for presentation).

Figure 3.4. *Standard Progressive Matrices Plus* 1996 Item-Equating Study
Item Difficulties (in Logits)
60 Items, Including ALL from Parallel Sets A and B and 5 Each from Parallel Sets C and D, Arranged in Order of Difficulty

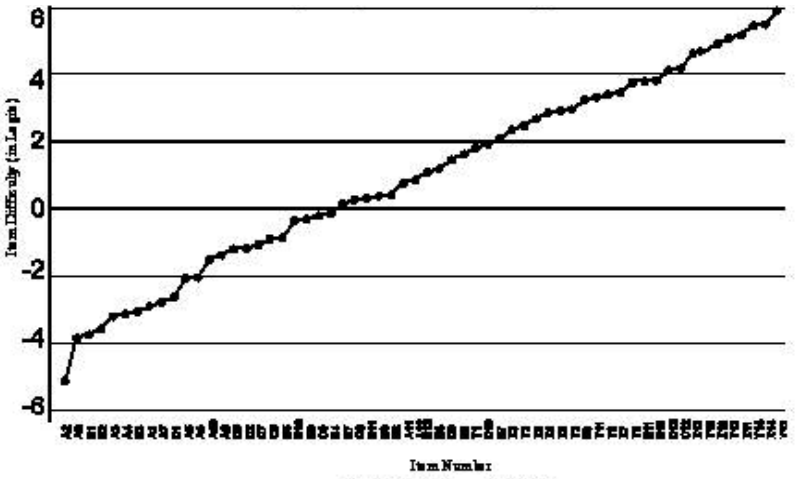
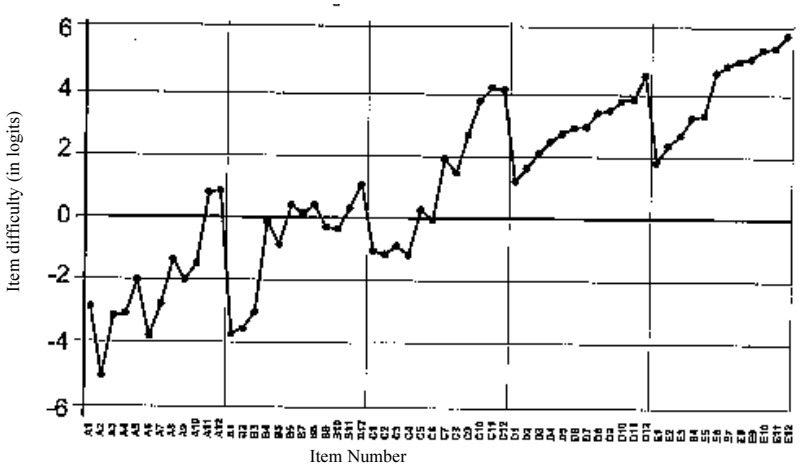


Figure3.5. *Standard Progressive Matrices Plus* 1996 Item-Equating Study
Item Difficulties (in Logits)
60 Items, Including ALL from Parallel Sets A and B and 5 Each from Parallel Sets C and D, Arranged in Sets



Appendix

The Design of Samples in Test Development

In other chapters of this book, attention is drawn to the need to have strictly representative samples of the populations to whom the results are to generalised if valid conclusions are to be drawn. More specifically, it is argued that representativeness is more important than size.

But, in test development, it is not only vital *not* to rely on random samples … *large* numbers are also required!

In the present study what was required was a design which would yield sufficient respondents with every score from the very lowest to very highest to make it possible to plot reliable Item Characteristic Curves (ICCs) for all the items.

The reasons for this are best illustrated via a hypothetical example, and coming at the problem from the other end. Let us start by making the (unrealistic) assumption that an equal number of people in a sample of 600 obtained each score from 1 to 60.

The ICCs show the percentage of those with each total score who get each item right. In the example we have chosen, there would be ten children having each score and it would be the percentage of each of these groups of ten which would be plotted to generate the ICCs.

Percentages calculated on bases of ten are obviously extremely unreliable. So, clearly, a much larger sample would be required to generate accurate data.

But, actually, if a random sample of the population had been tested, we would not in fact have got anything like equal numbers obtaining each total score from 1 to 60. Many would have scores around the average and there would be very few indeed having scores in the tails of the distribution, despite the fact that this is where most interest in testing lies. Consequently, the bases for the percentages of these low and high scores that got each item right (and which would which would be plotted to form the ICCs) would be very small indeed.

It follows from these considerations that, not only did we need to test far more than 600 people, we also needed to select our respondents in such a way that those obtaining both low and high scores were, by comparison with a random sample of the population, heavily over-represented. Put another way, an ideal distribution for our work would have been rectangular rather than bell-shaped.

In order to achieve something approaching this objective, we targeted three age groups which, it was hoped, would, between them, yield a significant number of people having each total score.

Having explored the merits of a number of designs, some of which would have required us to test very large numbers indeed, some of which were very cumbersome to administer, and others of which seemed likely to generate misleading information arising from fatigue or practice effects, the best compromise seemed to be that outlined in Table 3.1.

This design incorporated provision for checking the difficulties of the old items against the adjacent new items and, eventually, through retesting on the alternate form, direct checking of the difficulty indices of the new items against the old.

The design also enabled us to repackage the items into small subsets (booklets) so that information could be obtained from the same people on both old and new items without creating too great a burden in terms of time and fatigue.

In Table 3.1, O stands for Original Item and N for New Item. The numbers are the item numbers. Thus OA1 stands for Old Item A1, NA1 for New Item A1, and so on.

Readers who are contemplating work in this area may well find the account of the operational problems encountered in implementing the design sketched in Table 1 of interest and may therefore like to turn to Appendix 2 in the 1998-2004 edition of the SPM Section of the Manual[3.3] where these problems are described in some detail and credit given to those who helped surmount them.

Table 3.1. Sample Design for 1995 Item-Equating Study

Booklet Number	Target no.	Target to retest	Actual no. tested	Target age	Sets covered	Arrangements of sets	Total no. of items
Coloured Progressive Matrices							
1	150	25	287	5.5 - 8.5	A Ab B	OA1 NA2 OA3 NA4 ... OAb1 NAb2 ... OB11 NB12	36
2	150	25	274	"	"	NA1 OA2 NA3 OA4 ... NAb1 OAb2 ... NB11 OB12	36
3	150	25	373	"	"	NA1 - NB12	36
3B	150	25	164	"	"	OA1 - OB12	36
Standard Progressive Matrices							
4	150	25	238	11 12 13	A-E	OA1 NA2 OA3 NA4 ... OE1 NE2 ... OE11 NE12	60
5	150	25	240	"	"	NA1 OA2 NA3 OA4 ... NE1 OE2 ... NE11 OE12	60
6	150	25	224	"	"	NA1 - NE12	60
6B	150	25	215	"	"	OA1 - OE12	60
Restoration of discriminative power of Standard Progressive Matrices							
7	300		343	16 - 19	D E X	ND1 - NE12 + 1st 14 new items	38
8	300		267	"	D E Y	ND1 - NE12 + 2nd 14 new items	38

Notes

3.1. Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004)

3.2. Carver (1989)

3.3. Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004)

References

Carver, R. P. (1989). Measuring intellectual growth and decline. *Psychological Assessment, 1(3),* 175-180.

Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3: The Standard Progressive Matrices, Including the Parallel and Plus Versions*. San Antonio, TX: Harcourt Assessment.

Chapter 4

The Romanian Standardisation of the Standard Progressive Matrices *Plus*: Sample and General Results*

Anca Dobrean, John Raven, Mircea Comşa,
Camelia Rusu, & Robert Balazsi

Abstract

This chapter reports norms for Raven's Standard Progressive Matrices ***Plus*** (SPM ***Plus***) for Romanian people aged 6 to 80 years. A nationally representative sample of 2,801 people covering all demographic regions of the country was tested. The norms for the Romanian sample are slightly below recent norms from other countries. The test-retest reliability, assessed over a one-month period, was high and similar to that found in other countries.

Acknowledgements

We are grateful to the children and adults who took part in the study and also to the students in psychology who collected data. Without the willing cooperation of these thousands of individuals we would not have been able to report the study. We are grateful to Jean Raven who provided us help in data processing including preparation of the tables of comparative data. We would also like to thank colleagues from the Department of Psychology from Babes-Bolyai University, for their suggestions and support.

* An earlier version of this chapter has for some time been available on the Web Psych Empiricist

Introduction

The national standardization of a psychological assessment instrument represents a laborious task but is crucial to the effective use of assessment instruments. When this study was conducted there was no internationally accepted psychometric test with normative data for Romania. One reason for that is the subject of psychology, and with it psychological testing, was forbidden in Romania during the "communist" era. The discipline of psychology only obtained legal status in Romania in the winter of 1989.

Raven's *Standard Progressive Matrices* is one of the most widely used tests of general cognitive ability (MacKintosh, 1998). There are many reasons for this. The Raven Matrices are easy to administer, both as an individual and group test. The *Standard Progressive Matrices* (whether *Classic, Parallel,* or ***Plus*** version) consists of 5 sets of 12 matrices, gradually increasing in difficulty. The same test can be used for a wide age range. Also, the test has remained little changed from its original design (Penrose & Raven, 1936). The test format is non-verbal and can therefore be employed in diverse language cultures and in different settings (e.g., homes, schools, organizations). Viewed from the point of view of Spearman's (1927) ***g*** theory, Raven's Progressive Matrices are considered to be among the best measures of eductive ability (Bingham, Burke, & Murray, 1966). According to Raven, Raven, & Court (1998, revised 2003), the Raven Progressive Matrices (RPM) are among the best-established measures of human characteristics whose scientific status is most secure. Details of the construction and use of the *Standard Progressive Matrices* (SPM) and *Standard Progressive Matrices* ***Plus*** (SPM ***Plus***) will be found in Raven, Raven, & Court (2000 revised, up-dated, and extended 2004) and are summarised in other chapters of this book. The RPM are among the best predictors of academic and occupational performance and, especially, social mobility (Court & Raven, 1995).

One of the most widely discussed findings obtained with the RPM has been the so-called "Flynn Effect". James R. Flynn observed in the 1980's that the scores of different groups of people on standard intelligence tests had consistently increased over time. This effect was observed in many countries and especially on those measures - whether verbal or non-verbal - having the highest loadings on eductive ability (Flynn, 1987, 2000). Tests assessing mastery of traditional "academic" knowledge show much smaller increases. This secular trend in SPM scores emerges

quite clearly by comparing two standardizations of the SPM in the UK one from 1942 and the other from 1992 (Raven, et al., 1998 revised 2003; Raven, 2000). In both standardizations the test was administered to people from age 18 to 65. It would seem from the graphs published in these sources that the 50th percentile for 20 year olds rose from about 42 to 54 over this period. But it is obvious from these graphs that the latter figure underestimates the real increase owing to the test ceiling (maximum score 60). Over the whole century for which data are available, the 50th percentile rose from some 15 to 54. While it would appear that the lower percentiles rose more than the higher ones, it is immediately obvious from the graphs that this conclusion is invalid because the test ceiling has resulted in there being little discrimination above the 75th percentile from about 1950 onwards. In view of these results, research was put in hand to develop a test - which was eventually named the SPM ***Plus*** - which would restore the discriminative power at the upper end that the SPM had had when it was first developed. (See Raven et al. [2000 revised, updated, and extended 2004] for an account of the way in which this test was developed.)

The present study set out to build norms for SPM ***Plus*** for the Romanian population, to place those norms in the context of norms from other countries, and to report on the reliability of the test in Romania.

Method

Respondents

In order to maximize its representativeness, we built a three-stage stratified random sample. Stratifying a sample improves its representativeness by ensuring that various demographic groups are represented in their correct proportions. The participants in each strata are more homogenous on the stratification variables and also on the variables influenced by these. For example, if we stratify the population according to "type of location" we will obtain not only participants more similar on this variable but also on other variables related to this - such as education - which are, in turn, related to eductive ability (which is the subject of the study). The present sample was stratified by 18 cultural areas* and 7 locality types (3 rural

* They are simply cultural and historical units, not administrative ones. Counties from the same cultural area are similar in terms of urban development, rural development, land use patterns, ethnicity, and religious related culture (Sandu, 1999).

Table 4.1. *Standard Progressive Matrices* ***Plus***
Romanian Standardisation
Non-Contact and Refusal Rates

Reasons for non-contact and refusal	*Percent of target sample*
Refusals	17.03
Absence after 3 visits	8.19
Moved	0.73
Died	0.68
Institution at that address	2.68
Other	11.36

types and 4 urban types). This resulted in a total of 126 strata. Rural localities (communes) were grouped into 3 categories depending upon their level of development* (low, medium, high). Urban localities were grouped into 4 categories depending upon the number of inhabitants (under 30,000, 30,000 to 99,999, 100,000 to 200,000, and over 200,000 inhabitants). The three rural categories were low, medium, and high levels of economic development.

For each of the 126 strata we calculated the number of respondents to be tested in proportion to the number of people living in such areas in the total population of Romania. We kept only those strata that contained at least 10 participants (the cases from strata which had less then 10 cases were re-allocated to the most similar strata). For each layer we randomly extracted the localities from which we would select participants. In each locality we randomly selected two streets. For each of these streets we randomly selected every 5th house. In each house we selected everyone aged 6 to 80. Selection of houses and, of course, of respondents in a street ended when the required number of participants for that street had been obtained.

Because the development and decline of intellectual abilities is not constant across all ages from 6 to 80, increasing most rapidly between 6 and 17 years of age it was necessary to over-weight the population under 18 age in order to have enough respondents to generate adequate norms. In effect there were two samples, theoretically independent, for

* The development coefficient was based on indicators relating to the structure of the population, the demographic phenomena, modern goods endowment, possession of land and animals, and reported access to the nearest urban locality (Sandu, 1996, 1999).

6-17 and 18-80 year olds. Due to the combined effects of the random selection of houses (choice of people by selecting houses) and of the sample stratification procedure, we can only approximate the level of representativeness. Thus, we reached a maximal admitted theoretical error of ±2.8 (for the sample aged between 6 and 17) and ±2.5 (for the sample aged between 18 and 80) for p=0.05.

One of the chief problems confronting survey researchers is to gain the cooperation of the selected participants. The non-contact and refusal rates are presented in Table 4.1.

Such non-contact and refusal rates are typical for Romania. In order to check the quality of our sample we compared the demographic data for our sample with that for the population as a whole (Table 4.2). The national demographic data are based on the *Population and Dwelling Census, 2002*. The disproportionate number of young people was, of course, a deliberate result of having over-sampled them when selecting participants.

In summary, the sample had the following characteristics:

a. *Size*: 2,755 people aged 6 to 80, out of which 1,240 were aged between 6 and 17 (45%) and 1,535 aged between 18 and 80 (55%).
b. *Type*: three stage stratified random sample.
c. *Stratification criteria*: 18 cultural areas; size of the urban location (4 types), level of development of the rural locations (3 categories).
d. *Sampling*: random selection of locations (117 localities), places (199 streets), and houses. Within houses everyone aged between 6 and 80 was asked to take part.
e. *Representativeness*: the sample is representative for the non-institutionalized Romanian population aged between 6 and 80 with a ±2.5 % maximal tolerated error.

Procedure

The data were collected between October 2002 and November 2003. 178 interviewers carried out the testing. All of them were students of psychology specially instructed in how to apply the SPM ***Plus***. All the students had graduated from the psychological assessment course in the second year of studying psychology and also they followed a special training about the sampling procedure to be used. Their work was supervised and verified. The instructions for testing were taken from the SPM *Manual* (Raven et al., 2000 revised and extended 2004), translated

Table 4.2. *Standard Progressive Matrices* ***Plus***
Romanian Standardisation
Demographic Composition of Obtained Sample Compared with that of the Population (INSSE Data, 2002)

Variable	*Sample (%) Without weight*	*Population (%)*
Gender		
Male	44.9	48.0
Female	55.1	52.0
Age intervals		
5 - 9	8.8	5.7
10 - 14	23.5	8.1
15 - 29	27.9	25.5
30 - 49	22.3	29.7
50 - 59	9.3	11.2
Over 60	8.2	19.8
*Last school finished**		
Primary	4.9	19.4
General (8 classes)	15.6	26.2
Professional (10 classes)	20.3	17.6
High school	32.3	24.7
More than high school	24.3	12.1
Residence		
Urban	45.1	53.5
Rural	54.9	46.5
Ethnic group		
Romanian	90.3	89.5
Hungarian	7.1	6.6
Gipsy	1.3	2.5
Other	-	1.4
Car possession	-	38.0
TV possession	-	92.0

* Respondents over 18 years of age only.

into Romanian. All testing was carried out in the participants' homes and everyone was tested individually. Respondents were first asked to complete the SPM ***Plus*** and then to contribute socio-demographic data. After being instructed in the procedure to be followed, each participant worked individually and the administrator intervened only if the participant asked him/her to do so. Children, older people, and people who had difficulties completing the answer sheets were assisted by the interviewer. There was no time limit for the testing, but after 20 minutes a note was made of the item the person was working on at that time. The mean time taken to complete the SPM ***Plus*** was 43 minutes. The second step involved contributing socio-demographic data covering: occupational status, schooling, socio-economic status, and nationality. Participants had the option not to reveal this socio-demographic and private information (although confidentiality was assured).

Choice of test instrument

As mentioned earlier, the SPM, ***Plus*** was chosen because earlier studies had shown that it offered better discrimination among young adults with high cognitive abilities. Details of the construction of SPM ***Plus*** can be found in Raven et al. (2000, revised and up-dated 2004) and are summarised elsewhere in this book.

Results and Discussion

The test was scored in the usual way, assigning 1 point for a correct, and 0 for an incorrect, answer. All the analyses were conducted using raw, untransformed, scores. Respondents whose dates of birth or sex were missing were omitted from the analysis, as were those with severe neurological diseases. This reduced the sample size to 2,801.

Normative Data

Two considerations influenced the way we decided to present the normative data for the Romanian population:

1. Cross-sectional studies of the development and "decline" of eductive ability like our own, typically reveal a rapid increase until about 18 years of age and then what has, until recently, been interpreted as a steady decline into old age. In order to provide adequate reference data for those wishing to compare the scores

of individual young people and adolescents with normative data for their own age group it is therefore necessary for the norms for young people and adolescents to be presented at 6-monthly intervals. For adults more widely spaced - e.g., five-yearly - norms are adequate.

2. One way of assessing the significance of someone's score is to indicate the percentage of some appropriate reference group who obtain lower (or higher) scores. The advantages of this method are summarized in Raven at al. (2000 revised and extended 2004) but mention may be made of the following: (a) the non-Gaussian within-age group distributions invalidate the application of parametric statistics and the basis on which deviation IQs are calculated; (b) conversion of percentiles to deviation IQs with means of 100 and SDs of 15 exaggerates the discriminative power of the tests and thus lead users to place undue reliance on small differences in score; and (c) from the point of view of studying the development and decline of mental abilities, it is essential to study differential growth and decline among people of different levels of ability.

Smoothed summary norms for the Standard Progressive Matrices ***Plus*** (untimed) for Romania are presented in Table 4.3.

Inspection of Table 4.3 suggests that the effort made to restore the discrimination that the Classic SPM had at the upper end when it was first introduced without destroying its discrimination at the lower end was successful: The range of scores goes from a 5th percentile score for 6 ½ year olds of 8 correct answers to the 95th percentile for 18 year olds of 47. The latter leaves room for further increases in the scores of the top 5% of the population.

Comparative data

In addition to being standardized in Romania, the SPM ***Plus*** has also been, to some extent, standardized in the school district of Fort Bend, Texas, USA, and in Germany, Poland, and Hungary (see 2004 update of Raven 2000/2004 for further details). Table 4.4 presents a selection of the results. It will be seen that the Romanian norms, on the whole, lag behind those for the other countries. However, before concluding that the observed differences reflect genuine differences in eductive ability between countries it is necessary to consider the way in which the samples for the different countries were drawn. In the USA the sample (Raven

Table 4.3. *Standard Progressive Matrices* ***Plus***
Smoothed 2003 Norms for Romania

	Age in Years (Months)												
	6½	7	7½	8	8½	9	9½	10	10½	11	11½	12	12½
	6(3)	6(9)	7(3)	7(9)	8(3)	8(9)	9(3)	9(9)	10(3)	10(9)	11(3)	11(9)	12(3)
	to	to	to	to	to	to	to	to	to	to	to	to	to
Percentile	6(8)	7(2)	7(8)	8(2)	8(8)	9(2)	9(8)	10(2)	10(8)	11(2)	11(8)	12(2)	12(8)
95	22	24	28	31	32	33	34	35	36	37	38	38	39
90	20	22	26	30	30	31	32	33	33	34	35	36	37
75	19	20	22	25	27	28	28	29	29	30	31	32	33
50	**15**	**16**	**18**	**20**	**22**	**23**	**24**	**24**	**25**	**25**	**26**	**27**	**28**
25	10	11	12	14	15	16	17	18	19	20	21	21	22
10	9	9	11	11	12	13	13	14	14	14	15	16	16
5	8	9	9	10	10	10	11	11	12	12	13	13	14
n	20	30	28	31	37	22	31	44	51	62	55	66	73

cont.

Table 4.3. *Standard Progressive Matrices* ***Plus***
Smoothed 2003 Norms for Romania *(continued)*

	Age in Years (Months)											
	13	13½	14	14½	15	15½	16	16½	17	17½	18	18½
	12(9)	13(3)	13(9)	14(3)	14(9)	15(3)	15(9)	16(3)	16(9)	17(3)	17(9)	18(3)
	to	to	to	to	to	to	to	to	to	to	to	to
Percentile	13(2)	13(8)	14(2)	14(8)	15(2)	15(8)	16(2)	16(8)	17(2)	17(8)	18(2)	18(8)
95	39	40	40	41	42	44	45	45	46	47	47	48
90	38	39	39	40	40	41	42	43	44	45	45	46
75	34	34	35	36	36	37	38	39	40	40	40	41
50	**29**	**30**	**31**	**31**	**32**	**32**	**33**	**33**	**34**	**34**	**35**	**36**
25	23	23	24	24	25	25	26	26	27	28	28	29
10	17	18	18	19	19	20	21	22	22	22	22	22
5	15	15	16	16	17	17	18	18	18	18	18	18
n	72	80	77	80	62	58	51	56	64	63	40	45

cont.

Table 4.3. *Standard Progressive Matrices* ***Plus***
Smoothed 2003 Norms for Romania *(continued)*

	Age in Years (Months)											
	20	25	30	35	40	45	50	55	60	65	70	73+
	18	23	28	33	38	43	48	53	58	63	68	
	to	to	to	to	to	to	to	to	to	to	to	
Percentile	22	27	32	37	42	47	52	57	62	67	72	
95	49	48	47	46	45	44	42	41	40	38	36	34
90	46	45	44	43	42	41	40	39	37	34	32	30
75	42	41	39	38	37	36	34	32	30	29	27	25
50	**37**	**36**	**34**	**32**	**31**	**30**	**28**	**27**	**25**	**24**	**22**	**19**
25	29	27	25	24	22	21	20	19	17	16	15	13
10	21	20	18	17	16	15	14	14	13	12	12	11
5	16	15	14	13	12	12	12	11	11	10	10	10
n	158	157	142	157	148	188	150	123	85	74	58	63

Table 4.4. *Standard Progressive Matrices* ***Plus***
Smoothed 2003 Norms for Romania, in the Context of 1999 Norms for United States (Fort Bend), and 2000 Norms for Poland

	Age in Years (Months)									
	6 ½		8		12		15			
	6(3)		7(9)		11(9)		14(9)		14(6)	
	to		To		to		to		to	
	6(8)		8(2)		12(2)		15(2)		15(5)	
Percentile	**RO**	**FB**	**RO**	**FB**	**RO**	**FB**	**RO**	**FB**	**PL**	
95	22	30	31	36	38	41	42	46	49	
50	**15**	**18**	**20**	**25**	**27**	**33**	**32**	**37**	**39**	
5	8	7	10	12	13	23	17	29	30	
n	20	90	31	52	66	54	62	24	98	

	17			18		25		55	
	16(9)		16(6)	17(9)	17(6)	23	21	53	51
	to		to	to	to	to	to	to	to
	17(2)		17(5)	18(2)	18(5)	27	30	57	60
Percentile	**RO**	**FB**	**PL**	**RO**	**PL**	**RO**	**PL**	**RO**	**PL**
95	46	48	51	47	52	48	50	41	42
50	**34**	**39**	**41**	**35**	**42**	**36**	**39**	**17**	**29**
5	18	30	32	18	33	15	21	11	14
n	64	24	364	40	343	157	90	123	82

RO - Romania; FB - Fort Bend; PL - Poland

et al., 2000, updated 2004) was drawn from Fort Bend school district, Texas. This has a socio-economic level well above that typical of the US as a whole. The German adult sample (Raven et al., 2000, updated 2004) consisted mainly of voluntary participants and did not reflect the structure of the entire German population. The sample from Poland (Jaworowska & Szustrowa, 1993; Jaworowska & Szustrowa, 2000) was, so far as can be judged, representative, but it was based on a quota sample and it is well known that this procedure is subject to error because of its dependence on the discretion interviewers have in their selection of respondents who satisfy the specific socio-economic criteria that they are assigned. Given those facts, we cannot say whether the differences in performance between countries arise from basic differences in levels of eductive ability, from differences in the sampling procedures used, or from other reasons.

What is perhaps most revealing, however, is the comparison between the Romanian and the Hungarian results around the age of 18 presented in Table 4.5. In Hungary, by law, the entire age cohort who, in a particular year, becomes liable to perform military service must submit

Table 4.5. *Standard Progressive Matrices* ***Plus***
Smoothed 2003 Norms for Romania in the Context of Norms for Army Conscripts in Hungary and Army Recruits in Poland

	Age in Years (Months)				
	18	18½	18	18	20
	17(9)	18(3)			18
	To	to			to
	18(2)	18(8)			22
Percentile	**RO**	**RO**	**HU**	**PL**	**RO**
95	47	48	49	44	49
90	45	46	47	42	46
75	40	41	42	38	42
50	**35**	**36**	**37**	**34**	**37**
25	28	29	32	30	29
10	22	22	27	25	21
5	18	18	24	21	16
n	40	45	7,588	395	158

RO - Romania; HU - Hungary; PL - Poland

to psychological testing unless specifically granted exemption. The 1998 entry cohort all took the SPM ***Plus***. As can be seen from Table 4.5, the norms for the Romanian and Hungarian samples for this age group are similar.

It would seem from the above data taken together that, as one becomes more confident about the representativeness of the samples, the norms obtained become more and more similar to those obtained in Romania. And this conclusion is generally supported by the extensive collection of international data for the Classic *Standard Progressive Matrices* which have been brought together in Raven et al. (2000, revised and extended 2004).

Reliability analysis

Test-retest reliability

The test-retest reliability over an interval of one month, based on a sample of 142, 1st to 12th grade pupils was 0.88. The mean score at retest was 28.23. This compares with an original mean of 26.71, indicating a fairly typical increase in scores at re-test (Domuta, Comsa, Raven, Raven, Fischer, & Prieler, 2004; Domuta, Balazsi, Comsa, & Rusu, 2004). As a result, we can argue that the SPM ***Plus*** test will reflect with some accuracy the real performances of the test-taker both in the case of a single assessment (as in personnel recruitment) as well as in the case of repeated assessments of the eductive ability (as in the case of monitoring development).

Conclusions

The main objective of this paper was to describe the procedure adopted to standardize the SPM ***Plus*** in Romania and to discuss the results. We chose this form of the SPM because it seemed to have excellent discriminative power across the whole range of ability from early childhood, through adolescence, to old age. This hope and expectation was amply justified by the results. More specifically, it emerges that the construction of the SPM ***Plus*** *has* restored the discriminative power that the *Classic* SPM had at the upper end among adolescents and young adults when it was first developed but which had been eroded by the secular increase in scores that has come to be known as the "Flynn Effect". This increased

discriminative power has also been achieved in a way which leaves room for a continuing increase in scores. This improved discriminative power at the top end has been achieved without destroying either the test's discriminative power at the bottom end or its ability to register even the very lowest levels of performance.

So far as can be judged, the sample from whom the data were obtained is representative of the Romanian population and the norms based upon it are comparable with, and support, those established in other countries. The most likely explanation of the fact that the Romanian norms are slightly lower than those reported in most of the other studies is that the samples in these other studies left something to be desired.

But the similarity in the norms obtained in different countries is not the only thing that is impressive. The Item-Response-Theory-based internal consistency studies reported elsewhere (Domuta et al., 2004; Raven, Prieler, & Benesch, 2005) show that the tests' properties are remarkably stable across cultural groups. This, together with the test-retest correlation reported above, strongly suggest that the test offers a valid and reliable measure of eductive ability.

References

Bingham, W. C., Burke, H. R., & Murray, S. (1966). Raven's Progressive Matrices: Construct validity. *Journal of Psychology*, 62, 205-209.

Court, J. H. & Raven, J. (1995). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 7*: *Research and References*: *Summaries of Normative, Reliability, and Validity Studies and References to All Sections*. San Antonio, TX: Harcourt Assessment.

Domuta, A., Balazsi, R., Comşa, M., Rusu, C. (2004). Standardizarea pe populaţia României a testului Matrici Progresive Raven Standard *Plus*. *Psihologia resurselor umane,* 2(1), 50-57.

Domuta, A., Comşa, M., Balazsi, R., Porumb, M., & Rusu, C. (2003). Standardizarea pe populaţia României a testului Matrici Progresive Raven Standard *Plus*. În J. Raven, J.C. Raven, and J.H. Court: *Manual Raven*: *Sectiunea 3, Matrici Progresive Standard*, Editura ASCR, Cluj, 102-121

Domuta, A., Comşa, M., Raven, J., Raven C. J., Fischer, G., & Prieler, J. (2004). Appendix 4: The 2004 Romanian Standardisation and Cross-Validation of the Item Analysis of the SPM *Plus*. In J. Raven, J. C. Raven, & J. H. Court (2000, revised, updated, and extended 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3*: *The Standard Progressive Matrices Including the Parallel and Plus Version.* San Antonio, TX: Harcourt Assessment.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: what IQ tests really measure. *Psychological Bulletin,* 101, 171-191.

Flynn, J. R. (2000). IQ gains, WISC subtests and fluid ***g***: ***g*** theory and the relevance of Spearman's hypothesis to race. In G. R. Bock, J. A. Goode, & K. Webb (Eds.), *The Nature of Intelligence (Novartis Foundation Symposium 233)* pp. 202-227. Chichester, England: Wiley.

Institutul Naţional de Statistică (INSSE) (2002). Anuarul Statistic al României (Romanian Statistical Yearbook), Bucureşti.

Institutul Naţional de Statistică (INSSE) (2002). Recensământul populaţiei şi locuinţelor, Bucureşti.

Jaworowska, A., & Szustrowa, T. (1993). Polish Standardization of the RPM. *Polish Psychological Bulletin, 24*, 303-307.

Jaworowska, A., & Szustrowa T. (2000). *Podrêcznik do Testu Matryc Ravena*: *Wersje Standard, Standard Równoleg_a, Standard Plus. Polskie standaryzacje* (Raven Standard Progressive Matrices Manual: Classic, Parallel and Plus versions). Warszawa: Pracownia Testow Psychologicznych Polskiego Towarzystwa Psychologicznego.

MacKintosh, N. J. (1998). *IQ and Human Intelligence*. Oxford: Oxford University Press.

Penrose, L. S., & Raven, J. C. (1936). A new series of perceptual tests: Preliminary communication. *British Journal of Medical Psychology*, XVI, Part 2, 97-104.

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology*, 41(1), 1-48.

Raven, J. , Prieler, J, & Benesch. M. (2005). A replication and extension of the item-analysis of the Standard Progressive Matrices *Plus*, together with a comparison of the results of applying three variants of item response theory. http://wpe.info/papers_table.html

Raven, J., Raven, J. C., & Court, J. H. (1998, revised 2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1*: *General Overview*. San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (2000, revised, updated, and extended 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3*: *The Standard Progressive Matrices. Including the Parallel and Plus Version*. San Antonio, TX: Harcourt Assessment.

Raven, J. C. (1938/1962/2004). First published as *Progressive Matrices (1938).* London: H. K. Lewis. Revised and re-named as the *Standard Progressive Matrices* in 1962 with continuing publication by H. K. Lewis. Later published in Oxford, England, by OPP Ltd. and now by Harcourt Assessment, San Antonio, TX.

Sandu, D. (1996). *Sociologia tranziţiei. Valori şi tipuri sociale în România*, Staff, Bucureşti.

Sandu, D. (1999). *Spaţiul social al tranziţiei*, Polirom, Iaşi.

Spearman, C., (1927). *The abilities of man, their nature and measurement*. New York: Macmillan.

SPSS Inc. (1999). SPSS. Chicago: SPSS.

Chapter 5

Using the Romanian Data to Replicate the IRT-Based Item Analysis of the SPM+: Striking Achievements, Pitfalls, and Lessons*

John Raven, Joerg Prieler, and Michael Benesch

Abstract

In 2003 Raven's *Standard Progressive Matrices* ***Plus*** (SPM+) test was standardised on a nationally representative sample of 2,755 Romanians, aged 6 to 80. Using this data set it was possible to replicate and extend the Item Response Theory (IRT) based item analysis that had been conducted while developing the test. The correlation between the 1-parameter item difficulties (in logits) from the two studies was .96. More importantly, however, when the effects of applying different variants of IRT were compared, two striking conclusions emerged: (i) adoption of a one-parameter model - i.e. the most commonly employed variant of IRT - to data that really require a 3-parameter model can lead to seriously misleading conclusions. And, interestingly, as much or more can be learned by using the "unsophisticated" methods deployed by Raven in 1935 than by more recent statistical packages. (ii) The Figures displaying the Item Characteristic Curves for all 60 items of the SPM+ yield remarkable evidence of the scientific "existence" and scalability of Eductive (meaning-making) Ability. While these results are not new in an absolute sense, they will be new to many psychometricians, especially those steeped in classical test theory.

******

* An earlier version of this chapter has for some time been available on the Web Psych Empiricist: http://wpe.info/papers_table.html

This chapter has two objectives: (1) to report a replication and extension of the original item analysis of the *Standard Progressive Matrices* ***Plus*** (SPM+) test that was undertaken whilst the test was being developed, and (2) to report a study comparing the effects of fitting three variants of Item Response Theory (IRT) to the same data set.

An unexpected outcome of this research was a striking demonstration of the scientific "existence" and scalability of eductive (or meaning-making) ability - i.e. one of the two main components of Spearman's ***g***.

Although many of the conclusions from this work are not new in an absolute sense, they will be new to a wide range of psychologists and, indeed, to many involved in psychometrics, especially those steeped in Classical test theory.

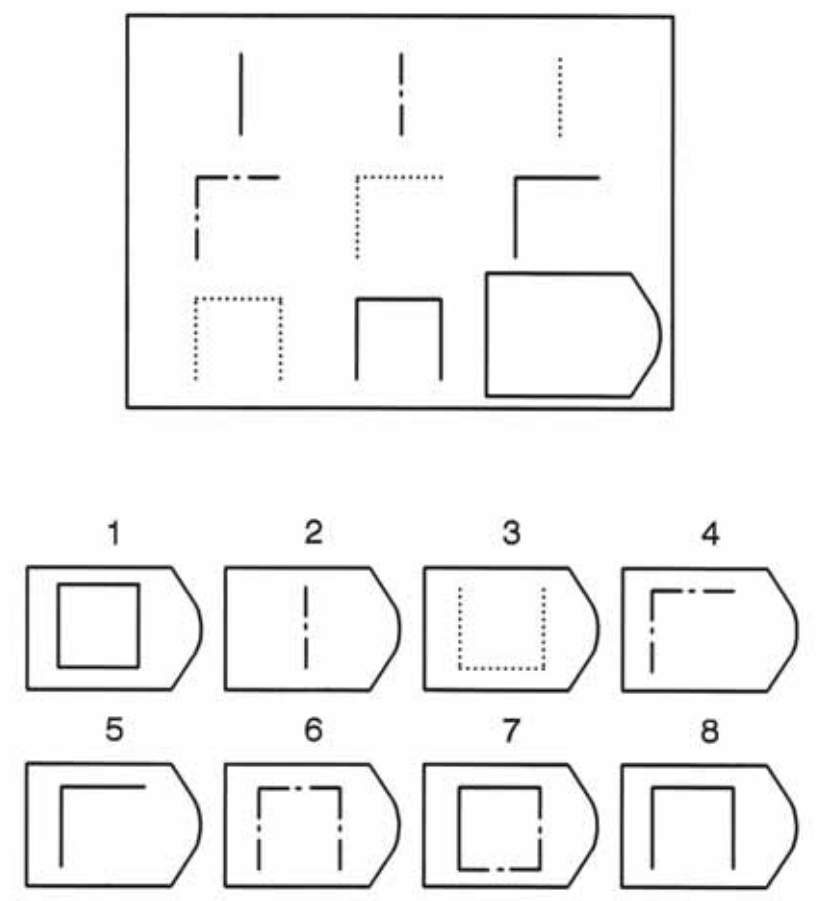
Background

As reported in the *General Introductory Chapter* to this book, Raven's *Progressive Matrices* are made up of non-verbal patterns, or designs, mostly with serial change in two directions. One part of the design is missing. Those taking the tests are asked to select from a number of options the part that is required to complete the design[5.1]. Figure 5.1 offers an illustration, although it is not an actual item from any of any of the tests.

The tests were developed to measure the *eductive* component of Spearman's ***g.*** In less technical terms, they were designed to measure the ability to make meaning out of confusion. It is generally agreed (see, for example, Carroll, 1997) that they do measure this ability. According to a survey carried out by Oakland (1995), Raven's *Progressive Matrices* tests are the second most widely used psychological tests in the world and a huge amount of fundamental research has been carried out using them.

The first form of the test was published in 1936. In order to distinguish it from other versions developed later this was re-named the *Standard Progressive Matrices* (SPM) in the late 1950s. The test was designed to facilitate the study of the development and decline of eductive ability from early childhood to old age and, in particular, for use in studies of the genetic and environmental determinants of variation in these abilities. For this reason, it was designed to discriminate across the entire range of mental ability and not to provide fine discrimination within any age or ability group. Particular care was taken to ensure that this discrimination

Figure 5.1. **An Illustrative Progressive Matrices Item**



would be achieved without creating frustration among the less able or fatigue or boredom among the more able.

In order to yield better discrimination among those of lower and higher ability, respectively, the *Coloured* and *Advanced* Progressive Matrices tests were later developed.

Nevertheless, at the time of its publication, the, 60-item, *Standard Progressive Matrices* (SPM) yielded excellent discrimination across the entire range of ability with the exception of less able older adults.

Unfortunately, as shown in particular by Flynn (1984a&b, 1987), Raven (1981, 2000b), Raven, Raven, and Court (1998, updated 2003), the scores achieved by samples of the general populations of many countries on the Raven Progressive Matrices (RPM) tests have

been increasing dramatically over the years. 50% of our grandparents would be assigned to Special Education classes in the US if they were evaluated against today's norms. [As an aside it is important to note that this increase has been documented on many measures of eductive (but not reproductive) ability, whether verbal or non verbal, for many other abilities (such as athletic ability), and for many other human characteristics such as height and life-expectancy. Readers interested in reviewing the evidence showing that these increases are not due to any of the obvious causes may find Raven, Raven, & Court (1998, updated 2004) and Raven (2000a&b) of interest.]

Because these increases eroded the ability of the SPM to discriminate among more able adolescents and young adults (among whom the test is widely employed) John Raven, Jnr., and his colleagues began, in the 1980s, trying to develop a new version of the SPM that would restore its ability to discriminate within these groups. The version of the test that finally emerged was named the *Standard Progressive Matrices* ***Plus.*** This is the test that we will be concerned with in this article.

The Measurement Model

Although it is well known that the items in the RPM tests become progressively more difficult (albeit in a cyclical format [which was introduced to provide training in the method of working]), it is not widely known that the *Standard Progressive Matrices* (SPM) was initially developed using a graphical version of what has since become known as "Item Response Theory". For example, in an article published in 1939, J. C. Raven included sets of graphs of the form that have since become known as Item Characteristic Curves (ICCs) for both the *Coloured* and *Standard* Progressive Matrices tests. These have been reproduced in the Introductory Chapter to this book. Similar graphs for the *Advanced* Progressive Matrices were included in the *Guide* to the use of that test which was published shortly after the Second World War (Raven, J. C., 1950). The Graphs in these articles (which correspond to those in Figure 5.4 below) plotted, for each item, the percentage of respondents with each total score who got the item right. The graphs for all items in the overall test (or the sub-set under investigation) were, as in Figure 5.4, included in a single plot so that they could be examined for cross-overs, spacing, and coverage of the domain of ability it was hoped to assess. The

objective was to select items whose curves had smooth ogives (instead of wandering all over the place), had ogives of approximately the same form, were equally spaced, and probed the whole domain of ability for which the test was intended. J.C. Raven argued that wandering ogives indicated that the items concerned were faulty. For example, there might be some feature of the item which confused more able respondents and distracted them from the correct answer. In a perfect world, the ogives would be vertical and equally spaced. One would then have the level of measurement achieved in a meter stick or foot-rule. There would be a 1 to 1 relationship between total score and final item passed.

Such an objective is not fully achievable in the measurement of human abilities so it is important, before moving on, to review a realistic analogy to illustrate what the measurement model is trying to achieve. The example taken is the measurement of the ability to make high jumps. When the bar is set low only the least able fail to clear it every time. Those who find it problematical do not always clear it and some members of this group clear it more often than others. So, even at a given height, the frequency with which it is cleared discriminates between the more and less able among those of a similar level of ability. In other words the graph of the percentage of trials in which it is cleared against total score increases steadily with overall ability. As the bar is raised, these curves, plotted on the same Figure, would follow one after the other across the page (see Figure 5.12 below). At a particular setting, the frequency of clearing the bar only discriminates among those of similar ability. By analogy, what one would wish to demonstrate if one set out to measure any psychological ability in a similar way would be that there is a systematic relationship between the Item Characteristic Curve for any particular item and the ICCs for all other items. These curves by definition show a systematic relationship between scores on any individual item and total score on the test (or statistically-based estimate of ability on the latent variable hypothetically being measured by the test).

There are several important lessons to be drawn out of this example:

1. The discriminative power of an item is given by the slope of the graph (Item Characteristic Curve, ICC) among those for whom the item is problematical. In other words, it is the correlation of the item with total score *within this group* (and not across the whole range of ability measured by the test) that indexes its discriminative power.

2. It would not make sense to try to establish the "unidimensionality" of the measure ("ability to make high jumps") by intercorrelating the "items" (centimetre marks on the post) across people (i.e. the accuracy with which information on whether they had cleared or failed to clear the bar at a particular level would enable one to predict whether they had cleared it at all other levels … i.e. calculating what would, in psychometrics be called the item-item correlations) and then either subjecting the resulting correlation matrix to factor analysis or calculating Alpha coefficients. The fact that someone clears the bar set at a low level will tell one very little about whether he or she will clear it at a high level so the correlation between the two will approach zero. Yet endless researchers steeped in classical measurement theory have done precisely this. That is, they have calculated and factor analysed the item-item correlations. This has led them to a host of entirely unjustifiable conclusions. For example, the fact that items of similar difficulty correlate with each other while the correlations between those items and items of very different difficulty are much lower has often been interpreted to mean that the RPM is not unidimensional but made up of items tapping a number of different "factors".
3. Introducing a time limit (e.g. what is the highest bar you can clear in 10 minutes, starting always with the lowest bar and running round in a circle in between) while still claiming that the test measures the ability to make high jumps creates utter conceptual confusion. Many of the most able will spend all their time running round in circles jumping over bars they can clear easily and never get a chance to demonstrate their prowess. Yet this is exactly what endless psychologists have achieved by administering the RPM, and especially the CPM and SPM (which pose the additional problem of a cyclical presentation designed to provide training and thus eliminate the effects of prior practice), with a time limit.

At this point we may draw attention to the way in which we have been using the term "Item Characteristic Curve". We are aware that some measurement theorists would like to restrict the term to graphs produced *after* transforming the data applying some mathematical variant of IRT (and, more specifically, plotting score on the *latent* variable being measured by the test, instead of raw score, on the horizontal axis[5.2]). However, as

we shall shortly show, such graphs typically render crucially important information invisible. 1-parameter models, for example, conceal what is happening to the proportions getting the item right before the item begins to be problematical to a significant proportion of those tested and differences in the slopes - discriminative power - item-total score correlations - of the items. To avoid confusion we have, in the remainder of this article, referred to the kind of graphs that Raven produced as "empirical" ICCs.

We turn now to the relationship between the graph-based variant of IRT developed by J. C. Raven in the 1930s and the mathematical variant developed by Rasch in the early 1950s (Rasch, 1960/1980)[5.3]. Rasch's task was to assess the long-term effects of a remedial reading programme from data collected in the course of a longitudinal study in which different tests had (necessarily) been taken by those concerned at different points in time as they aged (see the website referenced as Prieler & Raven, 2002 for a fuller discussion of the problems involved in measuring change). To do this, he had somehow to reduce the different tests to a common metric. To test the procedure he developed for the purpose, he applied it to the RPM and found that it worked (see Rasch, quoted by Wright in his forward to the 1980 edition of the previously mentioned book by Rasch). This fact is of greater significance than might at first sight appear in that an acrimonious debate has since raged around the question of whether the RPM "fits the Rasch model".

One question we wish to explore in this chapter is, therefore, what is lost (or gained) by fitting various mathematical variants of Item Response Theory to RPM data instead of plotting empirical ICCs.

The Development of the SPM *Plus*

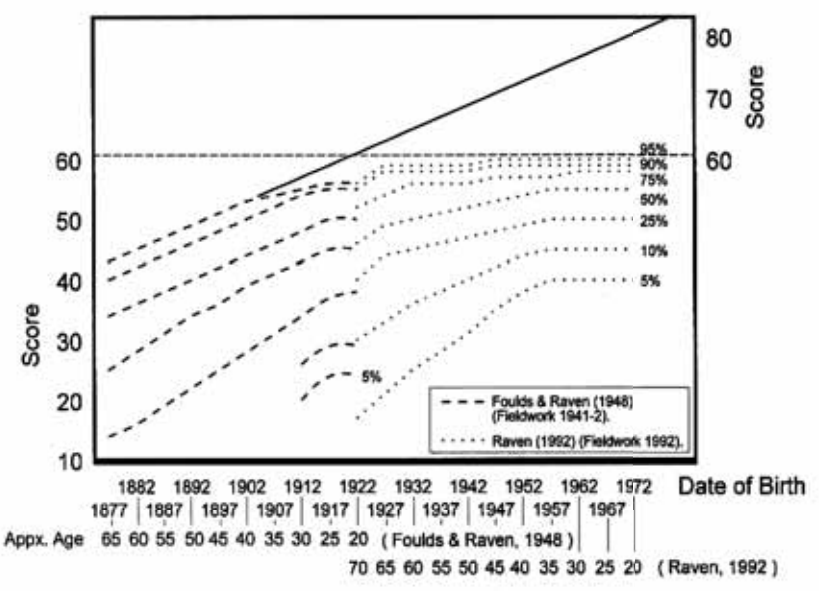
It turned out that the development of more difficult items for the SPM was no easy matter. Despite Vodegel-Matzen's (1994) outstanding work, it gradually became clear that there was much more to Raven's items than met the eye, and certainly a great deal more than Carpenter, Just, and Shell (1990) would have us believe. Indeed items generated for us by a widely cited authority on the rules governing the understanding and solution or Matrices items (who we will not name here) did not scale at all! The assistance of Irene Styles, Linda Vodegel-Matzen, and Michael Raven was therefore recruited. At first it was thought that the addition of

twelve more difficult items would be sufficient to restore the discriminative power that the SPM had had among more able respondents when it was first published. However, it gradually became clear that twice that number were required (see Figure 5.2, in which the graph showing the increase in the 95th percentile from those born in 1887 to 1912 has been extrapolated to 1982). Since we did not wish to modify the original SPM (for which such a vast pool of research data from so many countries existed), we also set about paralleling the existing items and checking that the proposed parallel items not only had equivalent difficulty to the old ones but also worked in the same way. To achieve these ends, a series of pilot studies of different sub-sets of old and new and more difficult items were undertaken. These were mostly conducted on about 300 respondents whose ages and ability levels seemed appropriate from the point of view of trialling the items concerned. The data from these studies were then analysed by Styles using 1-parameter mathematically-based IRT programs and the results used to whittle down the total pool of items to 108 that were carried forward into a full-scale item analysis. (The process is described in greater detail in Raven, Raven & Court, 2000/04.)

Assembling a sample that would enable us to conduct an adequate item analysis of the overall emerging test proved difficult indeed. Numerous researchers have come to entirely misleading conclusions about the scalability of the RPM as a result of not ensuring that their samples included sufficient respondents of all levels of ability. Under such circumstances it is obvious that certain items will fail to discriminate among those tested, will fail to correlate with total score, and will not take their "correct" place in the sequence of items. Even if a "random" sample of respondents of all ages and levels of ability were to be tested, there would, if the distribution was remotely Gaussian, be too few people in the tails to permit reliable item statistics to be calculated for the easiest and most difficult items.

But these were not the only obstacles. In addition to ensuring that we had enough low and high ability respondents to permit calculation of meaningful item statistics, we needed scope to discard items that were not working. In order to avoid widespread frustration (among younger or less able respondents) or boredom (among adolescents, adults, and more able respondents) and excessive testing times, it was therefore necessary to assemble a range of different booklets made up of items of differing difficulty with a view to later merging the data collected with different

Figure 5.2. *Classic Standard Progressive Matrices*
Implications of Score Increase for Revised Test Difficulty



booklets from different samples of respondents in the analysis.

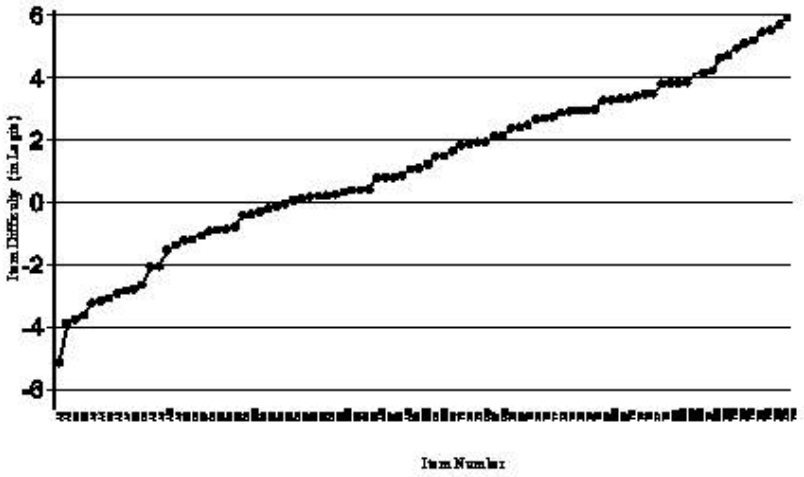
In the event, the testing of large numbers of young children was organised by Anita Zentai in Hungary, that of elementary and high school pupils by Rieneke Visser and Saskia Plum in the Netherlands, and that of University students by Linda Vodegel-Matzen in the Netherlands and Francis Van Dam and J. J. Deltour in Belgium.

The resulting data were again analysed by Styles using the previously mentioned statistical programs. At this point she unexpectedly encountered serious problems merging the various data sets, and was, in any case, restricted to 1-parameter Rasch analyses and unable to output sets of either IRT-based or "empirical" ICCs of the kind we had used in earlier studies.

We used the item-statistics she sent us to first reduce the total number of items from 108 to the 84 we thought we needed for the new test. However, as can be seen from Figure 5.3, a graph of the item difficulties of those 84 items revealed a number of plateaux (e.g. between items D3 and A11) where there were several items of similar difficulty. It was obvious that, if some of these items could be eliminated, we could re-create a 60 item test in which the ability of the *Classic* SPM to discriminate at

the upper end would have been restored without destroying its new-found ability to discriminate among the less able.

Figure 5.3. *Standard Progressive Matrices* ***Plus***
1996 Item-Equating Study
1-Parameter Rasch Item Difficulties (in Logits)
84 items - 60 Parallel Items and 24 Additional Items



It is also obvious from Figure 5.3 that it should be possible, when doing this, to achieve an almost linear relationship between the difficulty of the most difficult item that people were able to get right and their total score. Such a test would help to prevent certain researchers drawing inappropriate conclusions from their data. As Carver (1989) has shown, many researchers have discussed "spurts" in the development (and decline) of mental ability. Unfortunately, these often arise simply from a methodological artefact. It is obvious from Figure 5.3 that, as the most difficult items respondents are able to solve move across plateaux like those already mentioned, their raw scores increase with every new item they get right without there being a commensurate increase in the difficulty levels of the most difficult problems they are able to solve. A test having a linear relationship between total score and the most difficult item people were able to solve would eliminate this problem.

Unfortunately, from the point of view of eliminating items of similar difficulty, each of the Sets in the SPM (i.e. A, B, C, D and E) is made up of items of a different type. These not only require different forms of

reasoning but also introduce those being tested to the logic required to solve the next most difficult item in that Set. Elimination of the clearest candidates for removal would have resulted in a selection of 60 items which would have destroyed this unique property of the test. It would also have destroyed the comparability between the SPM and CPM. And it would have reduced the test's new-found ability to discriminate well among older adults and young children in zones where the 1938 version of the test did not work too well and which are of particular interest in the context of various Disabilities Acts.

As a compromise, the items making up Sets A and B in the new test were left intact. For the new Set C, five items were selected (on the basis of both item difficulty and an examination of their logic) to represent the logical stages of each of the old Sets C and D and supplemented by two new items.

The difficulty levels of the items which remained are plotted in Figure 5.5 below and, broken down by Set, in Figure 5.7.

The Romanian Study

In 2002/3 Domuta and her colleagues (Domuta, Comsa, Balazsi, Porumb, & Rusu, 2003; Domuta, Balazsi, Comsa, Rusu, 2004; Domuta, Raven, Comsa, Balazsi, & Rusu, 2004) standardised the SPM+ on a random sample of 2,755 Romanians, aged 6 to 80, tested individually in their own homes. The resulting normative data are compared with those from other studies in Domuta, Comsa, Raven, Raven, Fischer, & Prieler (2004).

Particularly because it covered such a wide range of ability, this study provided us with a superb opportunity to replicate and extend the item analysis that had been carried out whilst we were developing the SPM+ test. This was particularly important because, in that study, data relating to the items finally retained were collected when those items were presented to respondents in the context of different sub-sets of items, many of them of similar logic and difficulty drawn from the *Classic* SPM. Respondents' answers to the new items could well have been influenced by this context. The size and coverage of the Romanian sample not only goes a long way toward counteracting some of the problems known to be associated with calculating item statistics for the easier and more difficult items, it also meant that despite the, inherently unstable, nature of IRT-

based item statistics (Hambleton et al., 1991) there was a reasonable chance of obtaining meaningful data.

Sets of Item Characteristic Curves in the format originally published by Raven in 1939 and routinely published in the *Guides to the use of the RPM* in the '50s and '60, but this time generated by computer using a programme developed by Gerhard Fischer and applied to the data by Joerg Prieler are shown in Figure 5.4. Fischer's programme first applies a weighted normal "kernel smoother" to *every* subsequent set of 7 points to smooth the raw data and, in a second step, applies quadratic polynomials as 'splines' to draw a smooth curve through the smoothed points.

Graphing, Smoothing, and Transforming

At this point, a little more must be said about the graphing methods to be used to generate ICCs and, especially, the "empirical" ICCs. The original ICCs produced by Raven and his colleagues were drawn by hand after smoothing the raw data using the method of weighted moving averages. It is important to dwell for a moment on the reasons for this. As explained earlier, the individual graphs show, for each item, the proportion of those with each total score who got the item right. Given that scores on the SPM range from 5 to 60, only a few people in a random sample of the whole general population covering all ages from 5 to 90 will have high or low scores or fail the easiest items or get the most difficult items right. At these points one might therefore be talking about graphing percentages calculated on a base of 3 or 4 people. It follows that the points on which graphs are based in the "tails" of the ICCs for the easiest and most difficult items are particularly unreliable. It is therefore immediately obvious why it is necessary to smooth the data in some way - such as by the method of weighted moving averages - before plotting the graphs.

As computer programmes became more sophisticated, the smoothing was accomplished by fitting 4^{th} degree polynomials to the empirical data (see, for example, Graph RS1.10 in Raven, 1981). Unfortunately, one unanticipated consequence of the movement from mainframes to PCs turned out to be that, not only was it not possible - until Gerhard Fischer undertook the task - to find anyone who could reproduce the original (1935-1965) smoothing techniques by computer, we even lost contact with anyone who could easily generate graphs of the kind that had been produced by fitting 4^{th} degree polynomials to the data.

Figure 5.4. *Standard Progressive Matrices* ***Plus*** – Romanian Standardisation
Empirical Item Characteristic Curves for Items Comprising Sets A to E (Smoothed)

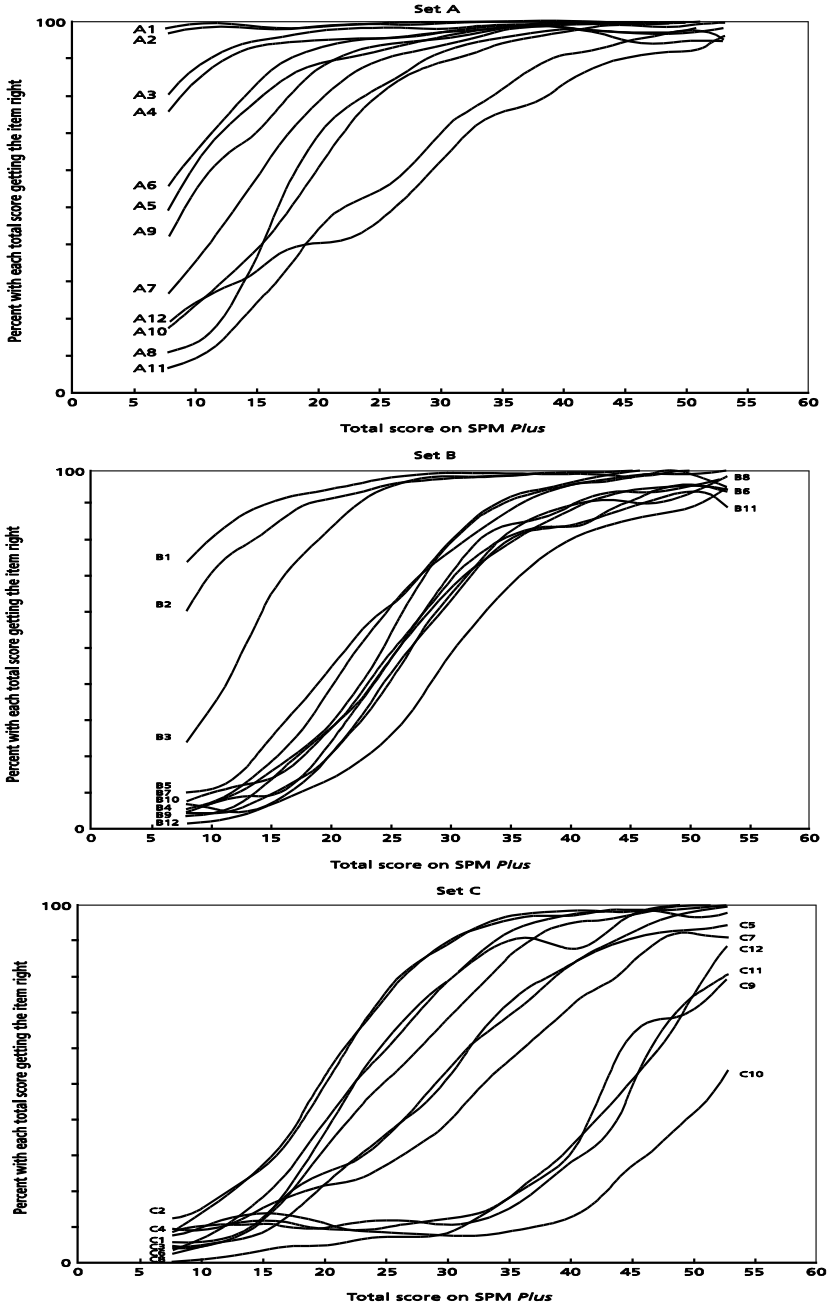
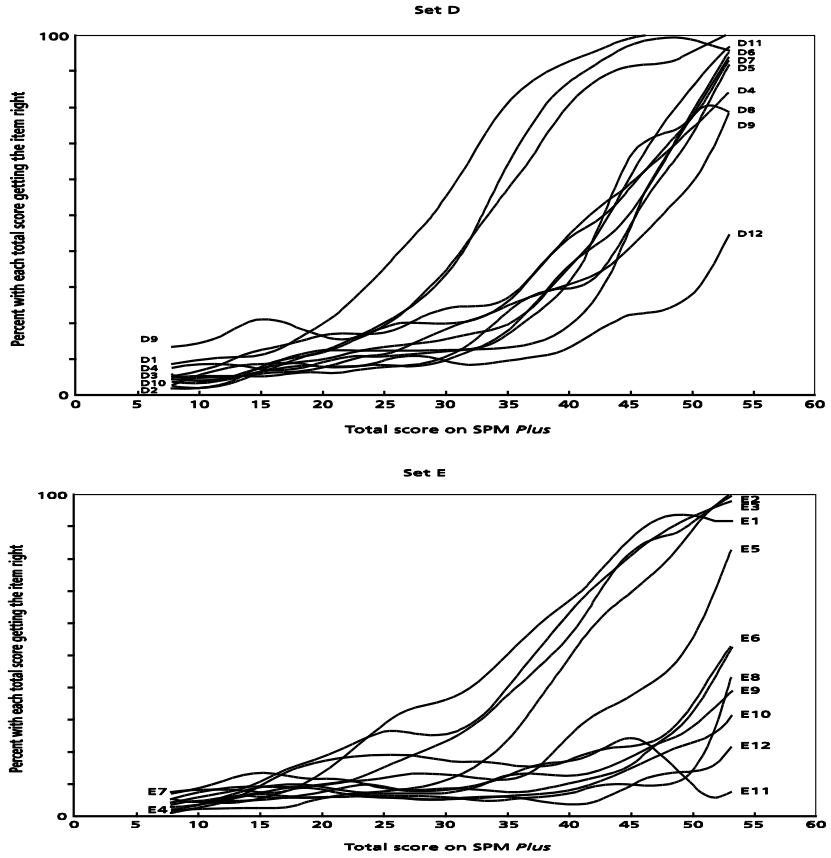


Figure 5.4 **Empirical Item Characteristic Curves for Items Comprising Sets A to E (Smoothed)** *(continued)*



The final straw that forced us to seek more vigorously for an alternative way forward was the discovery that Andrich and Styles were unable, even using their sophisticated RUMM programme, to plot more than 5 ICCs on a page (thus denying us the opportunity of studying cross-overs or the overall sequence and coverage of the items) or to fit the data with anything other than 1-parameter curves.

One point should perhaps be re-iterated here. Fischer's "reproduction" of the procedure originally employed when drawing the graphs by hand *smooths* the data. As we shall see, even the few mathematical-index oriented, computer based, IRT programmes that plot ICCs of the form published in Figure 2.4 in Hambleton, Swaminathan, and Rogers (1991)

transform the data on the basis of one variant of mathematical IRT (e.g., concerning the shapes and slopes of the graphs) *before* plotting them. The mathematical indices outputted by these programmes are also "contaminated" in exactly the same way. They fail to reveal the "raw truth" about the items. Thus they do *not* enable one to follow the recommendations the APA task force on statistical inference (APA, 1999), which encourage researchers to examine plots of their raw data before deploying "sophisticated" statistical programmes.

Some Implications of the Fischer-Prieler "Empirical" ICCs

Returning now to the Fischer-Prieler "empirical" graphs shown in Figure 5.4, attention may be drawn to the fact that a 3-parameter model is really required to fit these data. First, as can be seen most clearly in the graphs for Sets D and E, there is a clear "chance" or "guessing" component that results in a considerable number of people who lack the ability to solve many of the items logically choosing the correct answer "by chance". Second, although all the curves approximate the shape required by IRT, it is clear that they vary in slope. In other words, the effective correlation between the item and total score varies. Or, in still other words, the items vary in their discriminative power. Such variation counts against them in the most commonly employed mathematical version of IRT, which is the single-parameter Rasch model.

One reason why the single-parameter model is so widely used when a three-parameter variant is really required is that the latter is difficult to programme. But another is that, as Hambleton has perhaps emphasised more than others, IRT/Rasch indices, even for 1-parameter models, are unstable unless they are derived from very large data sets covering a wide range of abilities. These indices become even more unstable as two and three parameter models are fitted to the data. For these reasons - and because the computer programmes required to run 3-parameter models effectively are not readily available - most of the IRT-based statistics presented below are derived from the use of a 1-parameter model.

It is also apparent from the graphs for sets A, B, and C that, as will be seen more clearly below, the items are not as equally spaced as the graph of 1-parameter item difficulties derived from the item-equating and development study shown in Figure 5.5 below would lead one to expect.

1-Parameter IRT Analysis of data from Item Equating and Development Study Compared With 1-Parameter Analysis of Romanian Data

The correlation between the 1-parameter item difficulties established in the item-equating and development study and those emerging from the Romanian study was 0.96. It is therefore immediately obvious that the test properties are remarkably stable across populations and investigator.

Figure 5.5, reproduced from Raven et al. (2000, updated 2004), plots the non-recalculated item difficulties of the items retained in the final version of the SPM *Plus* after elimination of 24 items from the immediately preceding set. Figure 5.6 plots the corresponding data from the Romanian study but with the items arranged in the order of difficulty that emerged from that study. Again, a relatively straight line, with few plateaux or jumps, is obtained.

In Figure 5.7 the item difficulties from the Romanian study are plotted in the order in which the items appear in the published version of the test alongside the original plot from the item equating and development study (previously published as Figure SPM6 in Raven et al {2000 [ex 1998] updated 2004}).

The graphs for the original and Romanian data are strikingly similar. The relatively minor divergence among the more difficult items is due to the fact that the Romanian sample had too few people with high sores to permit the calculation of reliable item statistics. The irregular progression of item difficulty in Sets C and D in both studies is due to the compromises (summarised earlier) that had to be made in the selection and presentation of the items in the SPM+. The correspondence between these two graphs strongly confirms the inference that the properties of the SPM+ are remarkably stable across country, time, sample, investigator, and statistician.

A 3-Parameter IRT Analysis

After the above analyses had been completed, a way of running a 3-parameter analysis using the BILOG program was discovered. The resulting item statistics are shown in Table 5.1.

Two questions now arise:

1. How closely do the item difficulty indices calculated using the 3-parameter model correspond to those calculated using the 1-parameter model?

Figure 5.5. *Standard Progressive Matrices* ***Plus***
1996 Item-Equating Study
One-Parameter Item Difficulties (in Logits): 60 Items, Including ALL from Parallel Sets A and B and 5 each from Parallel Sets C and D, Arranged in Order of Difficulty

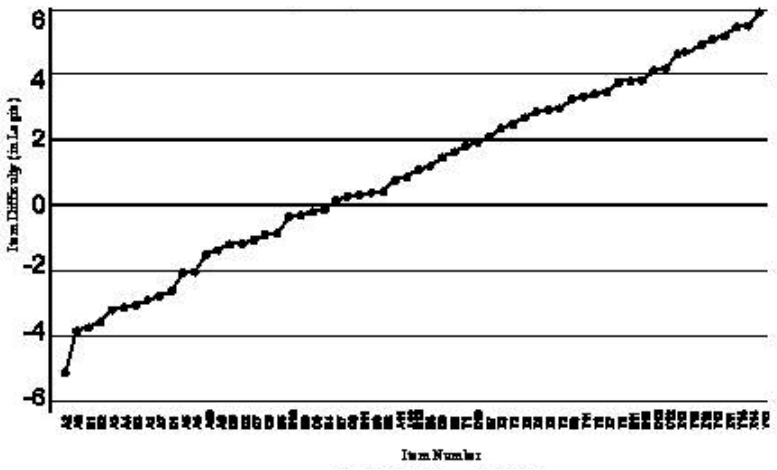


Figure 5.6. *Standard Progressive Matrices* ***Plus***
Romanian Standardisation
One-Parameter Model Item Difficulties Arranged in Order of Difficulty

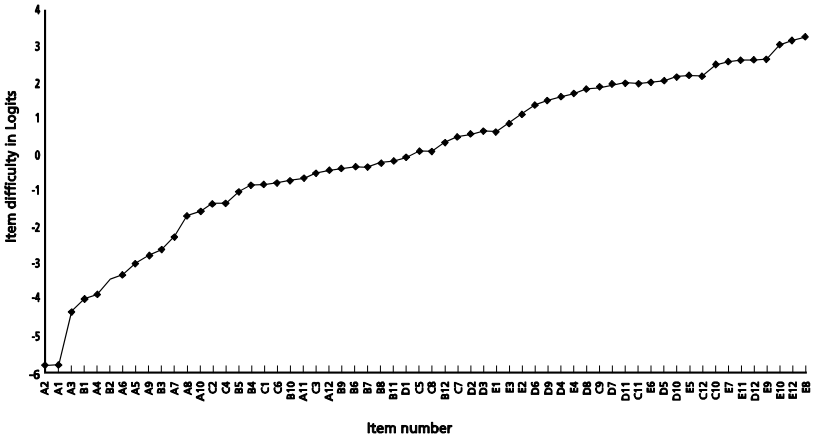
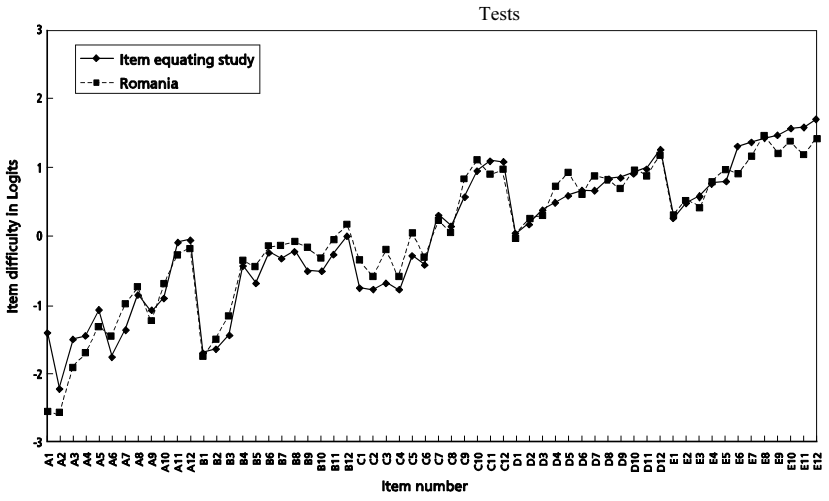


Figure 5.7. *Standard Progressive Matrices* ***Plus***
Comparison of Item Difficulties as Established in Item-Equating and Romanian Studies
One-Parameter Model Item Difficulties with Items Arranged in the Order in Which They Appear in the Tests



2. How much more, or less, information can be gleaned from looking at these indices than can be obtained by inspecting the "empirical" "ICCs"?

Given that we now had three sets of ICCs (the "empirical" ICCs, the ICCs derived from fitting a 1-parameter model, and those derived from fitting a 3-parameter model), it is possible to ask how much more closely the ICCs generated using a 3-parameter model correspond to the "raw" "empirical" "ICCs" than those generated using a 1-parameter model.

The item difficulties estimated using the LPCM Win (1998) and Winmira 1-parameter programs were identical. However, those generated using the BILOG 3-parameter programme were very different. Nevertheless, the correlation between the item difficulties derived from the 1- and 3-parameter models was 0.98.

We will shortly compare the information that can be extracted from the tables of 1- and 3-parameter item statistics with that which can be derived from looking at the empirical and other ICCs. But before doing so it is useful to compare the actual ICCs derived from the three models.

We first compared the "raw" or "empirical" ICCs generated by plotting 7-point moving, weighted, averages (as shown in Figure 5.4)

Table 5.1. *Standard Progressive Matrices* ***Plus***
Romanian Standardisation
3-Parameter IRT Item Statistics

	SETA			SETB			SETC			SETD			SETE		
Item.	Discrim	Diffic	Guess	Discrim	Diffic	Guess	Discrim	Diffic	Guess	Discrim	Diffic	Guess	Discrim	Diffic	Guess
.52	-5.73	0.13	0.96	-2.56	0.10	1.33	-0.40	0.03	1.44	0.21	0.12	0.89	0.74	0.07	0.07
.83	-4.03	0.13	0.91	-2.27	0.09	1.31	-0.66	0.09	1.69	0.54	0.08	1.10	1.10	0.09	0.09
.73	-3.27	0.12	1.27	-1.49	0.07	1.15	-0.18	0.05	1.23	0.61	0.06	1.22	1.08	0.15	0.15
.58	-3.40	0.12	1.15	-0.43	0.04	1.35	-0.61	0.12	0.98	1.51	0.08	1.61	1.27	0.07	0.07
.82	-2.04	0.11	1.21	-0.46	0.09	0.90	0.21	0.05	0.97	1.76	0.05	1.24	2.02	0.09	0.09
.93	-2.15	0.09	0.93	-0.15	0.03	1.02	-0.42	0.03	0.84	1.62	0.11	1.23	2.85	0.15	0.15
.91	-1.46	0.07	1.03	-0.07	0.07	0.87	0.71	0.11	1.31	1.59	0.07	2.35	2.38	0.10	0.10
.08	-1.00	0.04	1.20	-0.02	0.03	1.02	0.23	0.05	1.00	1.75	0.08	1.34	2.90	0.05	0.05
.94	-1.81	0.07	1.16	-0.14	0.03	1.31	1.69	0.10	1.16	2.17	0.18	0.91	3.11	0.08	0.08
.00	-0.94	0.07	1.52	-0.25	0.08	1.67	2.39	0.10	2.23	1.77	0.11	1.42	2.72	0.06	0.06
.83	-0.34	0.05	1.02	-0.01	0.03	1.79	1.74	0.11	1.62	1.51	0.07	0.47	4.14	0.06	0.06
.68	0.06	0.17	1.02	0.34	0.02	1.09	1.68	0.04	1.19	2.73	0.09	2.13	2.74	0.06	0.06

Notes:

Discrimination Index (Discrim.): When this is 0 it implies that the item does not discriminate between high and low scorers: the ICC is horizontal. An index of 1 indicates that the curve rises at 45 degrees. Larger numbers indicate a steeper curve.

Item Difficulties (Diffic.): These are analogous to Rasch logits. However, whereas Fischer's LPCMWIN calculates the item difficulties in such a way that their sum is always 0, this is not the case for the BILOG programme used here.

"Guessing" Index (Guess): This varies from 0 to 1, a 0 meaning that no "guessing" is taking place.

with those generated by fitting a 1-parameter model to the same data. This revealed that in certain cases, such as item C1 (Figure 5.8), the 1-parameter model curve seriously underestimated the discriminative power of the item - i.e. the "theoretical" curve was much flatter than the true curve. And, naturally, it failed to reveal the level of correct "guessing" occurring before respondents really possessed sufficient ability to set about solving an item correctly. These "guessing" levels varied from item to item and, in some cases, such as item D6 (Figure 5.9), showed a significant increase in the proportion of correct "guesses" that were made before the curve started to rise steeply.

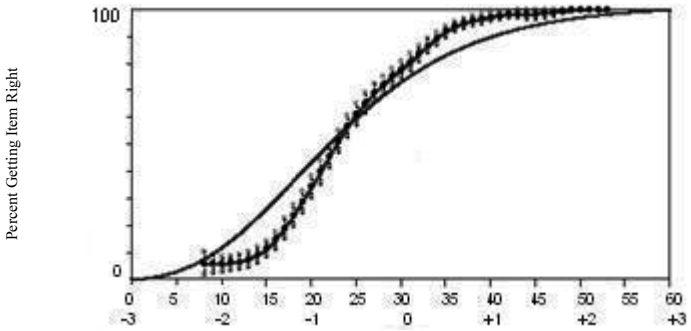
In Figure 5.10 the curve generated (with great difficulty) by fitting a 3-parameter model to the same data has been super-imposed onto the comparison of the empirical and 1-parameter ICCs for item C1 shown in Figure 5.8. Figure 5.11 presents similar comparative curves for Item D6. It will be seen that, in both cases, the curves generated by the 3-parameter model fit the data almost perfectly.

We turn now to summarising what, it seems to us, can be learned by comparing the *indices* of discriminative power derived from a 3-parameter model shown in Table 5.1 with what can be learned from studying the empirical ICCs and those generated using the 1 and 3-parameter models (only two samples of which have been reproduced above in Figures 5.8, 5.9, 5.10, and 5.11). It is abundantly clear that the variation in the 3-parameter discrimination indices does indeed reflect the observable variance in the slope of the empirical curves and generally does a much better job of reflecting the item characteristics than the graphs derived from fitting a 1-parameter model to the data.

Nevertheless, all was not quite assured. For example, when we compared what could be learned from looking at the "empirical" ICCs for items D9 and D10 in Figure 5.4 with what the mathematical indices appeared to be telling us, we found that, yes, D10 does indeed have better discriminative power than D9, but, no, D9 is *not* more difficult than D10 as the 3-parameter indices suggest.

We may focus now on the question of "guessing". However, by way of introduction, it is useful to draw attention to the fact that we have shown elsewhere (e.g. in the Addendum to Raven et al., 1998, updated 2004) (and our work has been confirmed by such authors as Carpenter, Just, & Shell, 1990, Vodegel Matzen, 1994, and Hambleton et al., 1991), that the term is a misnomer because, when an item is too difficult for people, they do not usually choose their answers at random but are guided by an hypothesis, albeit an incorrect one.

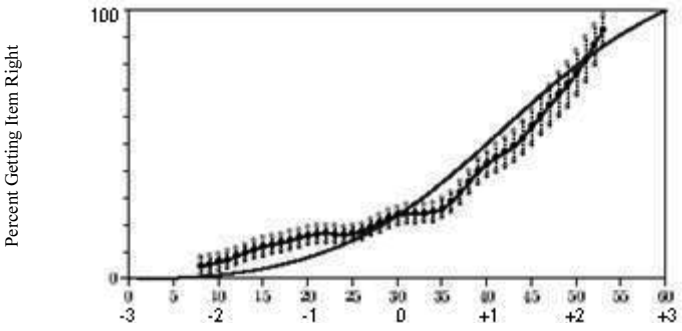
Figure 5.8. *Standard Progressive Matrices* ***Plus***
Romanian Standardisation
Comparison of Empirical and 1-Parameter ICCs for Item C1



Empirical curve (starred line): Number of items answered correctly, i.e. total score on test.
*1-Parameter IRT curve (solid line): Ability estimates in logits**

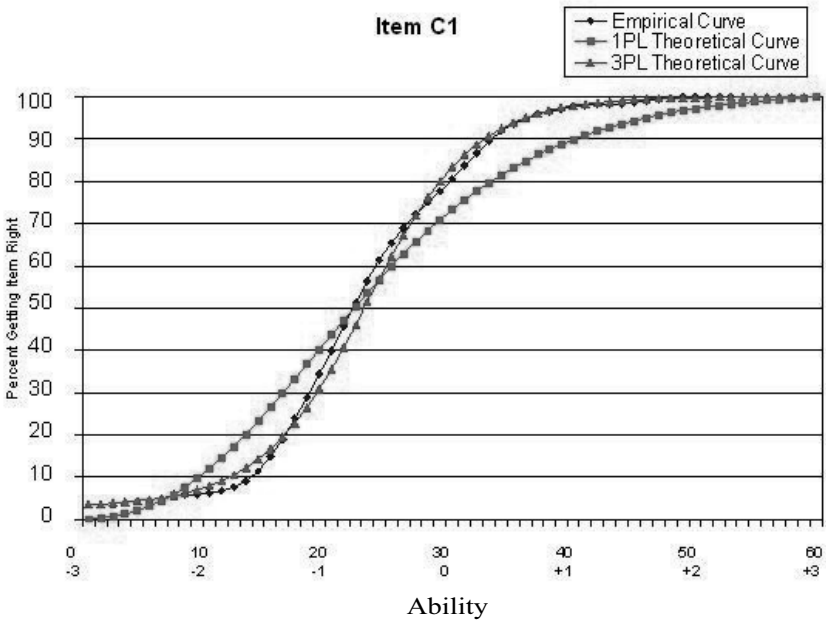
*IRT ability estimates are calculated from the difficulty and discriminative power of all items operational at each ability level.

Figure 5.9. *Standard Progressive Matrices* ***Plus***
Romanian Standardisation
Comparison of Empirical and 1-Parameter ICCs for Item D6



Empirical curve (starred line): Number of items answered correctly, i.e. total score on test.
1-Parameter IRT curve (solid line): Ability estimates in logits

Figure 5.10. *Standard Progressive Matrices* ***Plus***
Romanian Standardisation
Comparison of Empirical, 1-Parameter, and 3-Parameter ICCs for Item C1

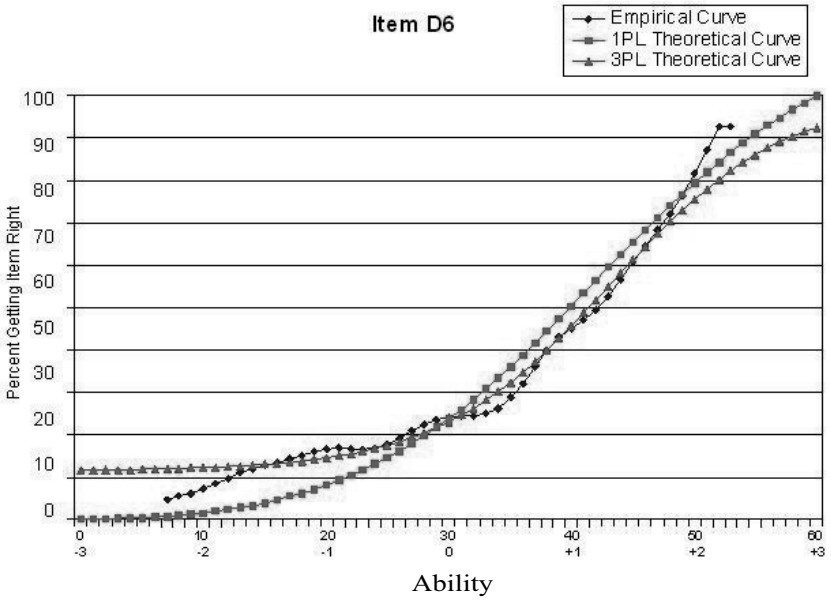


Empirical curve: Number of items answered correctly, i.e. total score on test.
1- and 3-Parameter IRT curves: Ability estimates in logits

The individual ICCs shown in Figure 5.4 for items D4, D5, and D6 may first be compared with each other and with that for E3. For items D4 and D5, "guessing" is clearly occurring, but the level is below that to be expected by chance and remains constant. For D6, the level is again constant, but higher. Both of these effects are reflected in the guessing statistics in Table 5.1, although one might be tempted to think that the very low figures for D4 and D5 mean there is no guessing going on. In fact there *is* "guessing" going on but it is below the level expected by chance.

However, if we turn to the ICC for item E3, we can see from Figure 5.4 that a considerable number of people seem somehow to be getting this item right before they have the level of ability that seems to be required to solve it correctly. This is reflected in the "guessing" statistic for this item in Table 5.1. Both observations suggest that it might be possible to improve the discriminative power of the items by tinkering with the distracters.

Figure 5.11. *Standard Progressive Matrices* ***Plus***
Romanian Standardisation
Comparison of Empirical, 1-Parameter, and 3-Parameter ICCs for Item D6



Empirical curve: Number of items answered correctly, i.e. total score on test.
1- and 3-Parameter IRT curves: Ability estimates in logits

Comparison of plots of all 60 ICCs derived from 1pl and 3pl models

Figures 5.12 and 5.13 show the plots of the ICCs of all items derived from the 1 and 3-parameter models. Nothing could give a better impression of the difference between the conclusions that follow from forcing the data into these alternative models. When the data are forced into a 1-parameter model, the ICCs appear to be of the same shape and evenly spaced. This is, presumably, a result of having employed item statistics derived from fitting a 1-parameter model to the data collected in the course of the item-equating and development study to select the items that were actually retained in the test. But the plot of the 3-parameter curves look very different indeed. The items are not equally discriminating; the order of difficulty varies with the ability of those taking the test, the items are not equally spaced, and they do not probe the domain of ability

to be sampled by the test anything like as well as the plot of 1-parameter ICCs would have us believe. It is almost certain that, had we had sets of 3-parameter graphs generated from the data collected during the item-equating and development study we would have modified item A12 -- which is the item whose ICC crosses those for all the other items in the test. It is apparent that many of the least able respondents get it right before they have the ability to solve it and quite a number of the most able still get it wrong. Clearly, something about the item is distracting these able respondents.

Although these may appear to be relatively minor quibbles here, it is important to recall that, before the final version of the SPM+ test used in the large Romanian study (from which the data used here was drawn) was published, its items had been extensively worked over. Many had been rejected and others revised. More striking evidence of what can be learned from viewing sets of 3-PL ICCs can be found in Figure 5.9 in the next chapter (in which we report a the results of a pilot analysis of data collected in the course of developing a Romanian version of the Mill Hill Vocabulary test).

Although it is not possible in that Figure to identify which curve belongs to which item, it is obvious from their ICCs that some of the items are functioning very poorly: their ICCs cross those for *all* the other items. Far too many low ability people get these items right and far too many high ability people never get them right. In other words, there is something about these items which leads low ability people to select the correct answer and something which distracts the more able from doing so. (As will be seen from other material presented in that chapter, examination of the Item Distracter Curves enables us to be clearer about what, exactly, the problem is.)

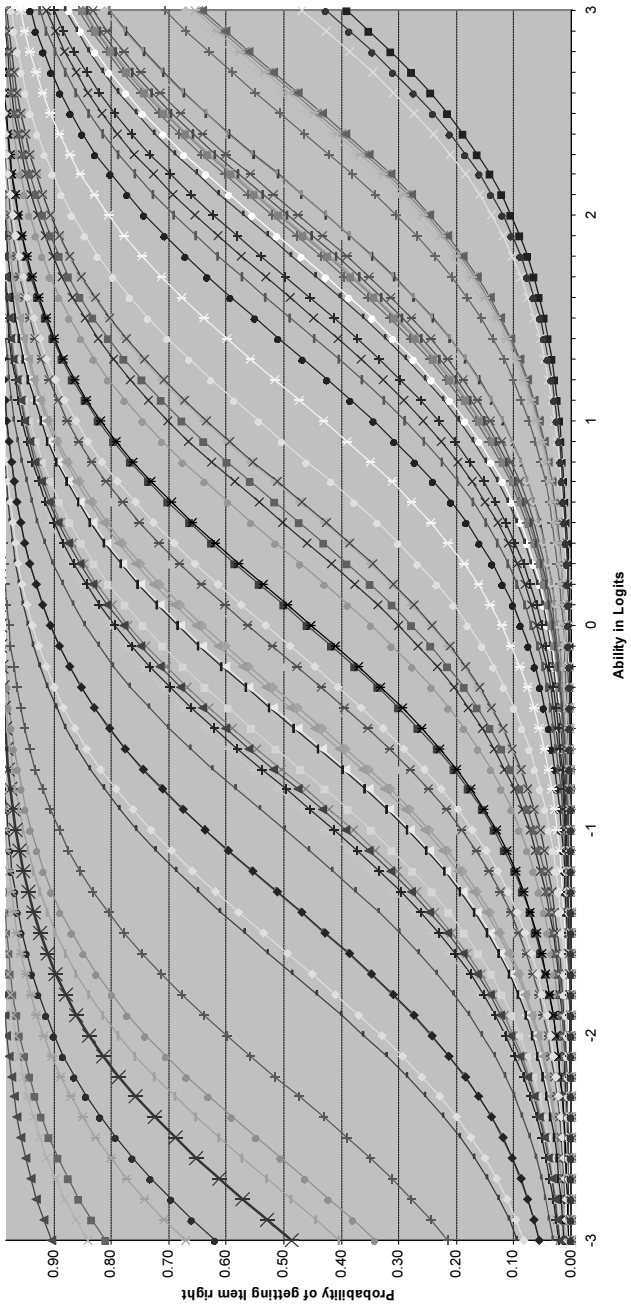


Figure 5.12. *Standard Progressive Matrices* ***Plus***
Romanian Standardisation
1-Parameter Model Item Characteristic Curves
(Each graph represents one Item)

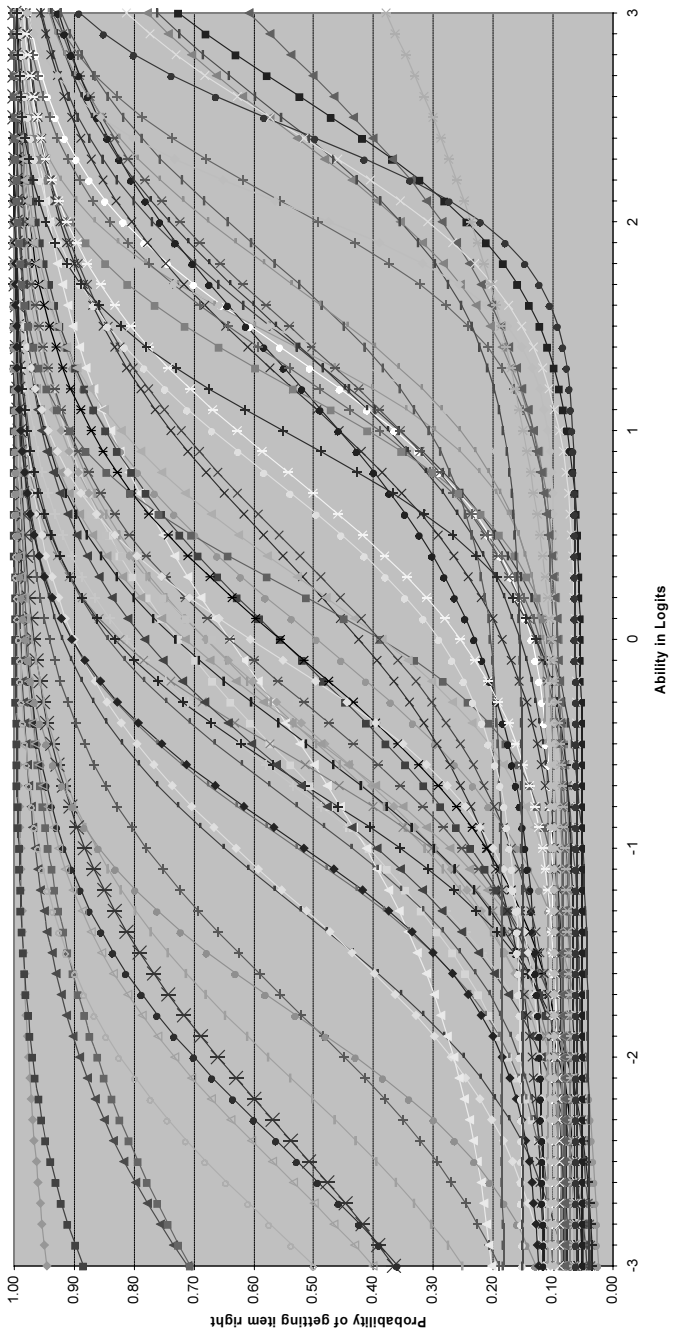
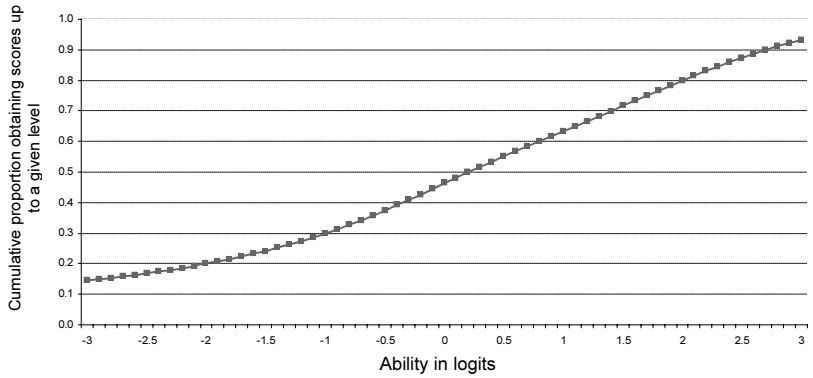


Figure 5.13. *Standard Progressive Matrices* ***Plus***
Romanian Standardisation
3-Parameter Model Item Characteristic Curves
(Each graph represents one Item)

Test Characteristic Curve and Test Information Function Curves

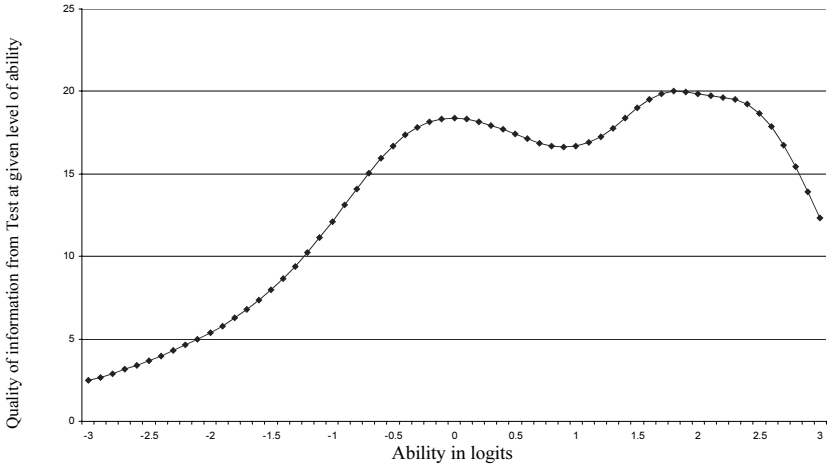
Figure 5.14. *Standard Progressive Matrices* ***Plus***
Romanian Standardisation
Test Characteristic Curve for 3-PL



The Test Characteristic Curve and Test Information Function curves are displayed in Figures 5.14 and 5.15. The former shows how total score on the test varies with Ability, assessed in logits. Thus if there were, as shown earlier, areas in which there were many test items of similar difficulty, the total score on the test (vertical axis) would increase steeply but not be reflected in much change in ability. The Test Information Function curve shows how much differential information the test yields at different points in the scale. Thus, if, as is commonly the case, the Test Information Function (TIF) curve approximates a Gaussian curve, it means that the test discriminates well among those with moderate levels of ability and does a poor job among among those with high or low ability. If one of the uses of the test is, for example, to differentiate among those who have been referred as potential candidates for Special or Gifted education, this is not exactly desirable. Thus, contrary to what might be expected, the ideal shape for a test information function curve might be rectangular or even bimodal. (See Hambleton *et al,* 1991 for a fuller discussion).

It will be seen from Figure 5.14 that the Test Characteristic Curve bears a marked resemblance to the approximately straight line of item difficulties shown in Figure 5.5. And, from Figure 5.15, it will be seen that the Test Information Function curve bears a marked resemblance to the overall plot of the distribution of raw scores shown in Figure 5.8 in

Figure 5.15. *Standard Progressive Matrices* ***Plus***
Romanian Standardisation
Test Information Function



the chapter by Prieler & Raven on the *Measurement of Change*. Many readers will expect both the overall distribution of raw scores and this TIF to approximate to a Gaussian (often believed to be a "normal") distribution and therefore believe that it shows that they show that there is something wrong with the test. Nothing could be further from the truth. The appropriate shape for TIFs has just been discussed. But a comment may also be made concerning expectations re the distribution of raw scores. Let us suppose for a moment that that the *within-age* distributions -- i.e. those used to determine percentile scores (themselves often converted to deviation IQ scores) -- were Gaussian (which they are not), then it would not be possible for the overall distribution, combining all age groups, to be Gaussian.

More importantly, a Gaussian TIF would imply that the test had best discrimination around the mean and poor discrimination in the tails. And, indeed, this is the case for most tests. But what one actually wants is, at least, equally good discrimination across the entire range of ability for which the test is intended, and, perhaps, superior discrimination in the tails … i.e. around the 5^{th} and 95^{th} percentiles -- these being the values at which most tests are applied. Thus the ideal TIF curve would be rectangular, even bi-modal. And this is, of course, exactly what Figure 5.15 suggests that the SPM+ comes close to delivering! (Hambleton et al., 1991, include a powerful discussion of this point.)

Some Conclusions

One conclusion which may be drawn from this exercise seems to be that a careful examination of the empirical ICCs in general tells us more than an examination of the sophisticated IRT item parameters ... but the IRT item parameters may lead us to pay more attention to the details of what the ICCs are telling us!

A still more striking conclusion is that an all-item plot of the 3-parameter ICCs very quickly gives what seems to be a fairly accurate impression of which items are working "correctly" and which merit more attention.

Although it was not among the topics we set out to explore in this study, it would seem that we have accidentally stumbled upon a striking demonstration of the scientific "existence" and scalability of eductive ability. Although Figure 5.12 perhaps supports this impression somewhat too strongly (although it is but a graphical version of the model that is most widely used), it is obvious from Figure 5.13 that more judicious work on the items could result in a test which did, in reality, have the properties suggested by Figure 5.12. However, having said that, it is perhaps important to caution that measurability in no way implies a single underlying causation. Although the hardness of substances can be scaled in exactly the same way as the eductive ability of human beings that in no way implies that the variation in hardness between substances is due to any single underlying factor. Further, related, points can be drawn out of the analogy with measuring the ability to make high jumps. No one would claim that high jumping ability was determined by a single underlying ability in the way in which the scaleability of the RPM is often used to justify the inference that the variance is determined by a single underlying ability. Nor would they seek single-variable explanations of the increase in the ability over time. Nor would they argue that, because there are no more Olympic medallists the general increase in the ability over time is unreal. Nor would they claim that the fact that training can increase the ability invalidates the theoretical concept being measured. And nor would they back-project the increases in high jumping ability over the past century to the time of Greeks and argue that, since the Greeks were demonstrably not such poor athletes, this means that our measure of high jumping ability is invalid.

Notes

5.1. As we will shortly see, many researchers, as a result of not appreciating the psychometric model deployed in the development of the tests, have drawn inappropriate conclusions from their research. Similar errors have arisen from failure to appreciate why the designs have been termed "matrices". At root, the word "matrix" refers to self-sustaining progressive development as in the womb. The next step in development is determined by the multi-dimensional pattern that has already been established. If the step that is actually taken does not conform to the emerging pattern, one has an abortion, or at least a deformity. It is this usage that lies at the heart of the way the term is used in mathematics - and, indeed, all the items in the RPM are capable of being expressed as mathematical matrices. Confusingly, however, the term "matrix" is also widely used to refer to any rectilinear array of words or data irrespective of whether it has an internal progression or order.

5.2. It has been put to us that, if the data fitted a 1-parameter model (which they conspicuously do not) the raw score might be treated as a reasonable approximation to score on the latent variable but that this assumption cannot be made if a 3-parameter model is required. Unfortunately, as we will see later, there are much more serious reasons why, even if the data fitted a 1 parameter model, this approximation should not be accepted. The fact remains, however, that few researchers, even if they adopt IRT instead of classical test theory, go to the trouble of fitting the right model, never mind transforming their data to scores on latent variables before conducting their analyses

5.3. Other psychologists who anticipated, or contributed to, the development of the variant of IRT that became the dominant model, largely as a result of being popularised by Wright in the 1960s, include Guttman (1941), Lawley (1943), Lazarsfeld (1950), and Lord (1952).

References

APA Task force on Statistical Inference. (1999). See: L. Wilkinson and Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97(3)*, 404-431.

Carroll, J. B (1997). Psychometrics, intelligence, and public perception. *Intelligence, 24,* 25-52.

Carver, R. P. (1989). Measuring intellectual growth and decline. *Psychological Assessment, 1(3),* 175-180.

Domuta, A., Balazsi, R., Comsa, M., & Rusu, C. (2004). Standardizarea pe populaţia României a testului Matrici Progresive Raven Standard *Plus*. *Psihologia resurselor umane,* vol. 2, nr. 1, 50-57.

Domuta, A., Comsa, M., Balazsi, R., Porumb, M., & Rusu, C. (2003). Standardizarea pe populatia Romaniei a testului Matrici Progresive Raven Standard *Plus*. In J. Raven, J. C. Raven, & J. H. Court: *Manual Raven*: *Sectiunea 3, Matrici Progresive Standard*, Editura ASCR, Cluj, 102-121.

Domuta, A., Comsa, M., Raven, J., Raven C. J., Fischer, G., & Prieler, J. (2004) Appendix 4: The 2003 Romanian Standardisation and Cross-Validation of the Item Analysis of the SPM *Plus.* In Raven, J., Raven, J. C., & Court, J. H. (2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3*: *The Standard Progressive Matrices Including the Parallel and Plus Version*.

Domuta, A., Raven, J., Comsa, M., Balazsi, R., & Rusu, C. (2004, submitted for publication). The Romanian Standardization of Raven's Standard Progressive Matrices *Plus.*

Fischer, G. H., & Ponocny-Seliger, E. (1998). *Structural Rasch Modeling.* Handbook of the Usage of LPCM-Win 1.0, Groningen: ProGAMMA (www.scienceplus.nl).

Flynn, J. R. (1984a). IQ gains and the Binet decrements. *Journal of Educational Measurement, 21*, 283-290.

Flynn, J. R. (1984b). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29-51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171-191.

Guttmann, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In: Horst, P. et al. (Ed.): *The prediction of personal adjustment*. New York: Social Science Research Council.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press.

Lawley, D.N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, 61, 273-287

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In: Stouffer, S. A. et al. (Ed.): *Studies in social psychology in World War II, No. IV. Measurement and Prediction.* Princeton: Princeton University Press, 362-412

Lord, F. M. (1952). *A theory of test scores*. Psychometric Monograph, No. 7. Iowa City, IA: The Psychometric Society.

McKinzey, R. K., Prieler, J. A., & Raven, J. (2003). Detection of children's malingering on Raven's Standard Progressive Matrices. *British Journal of Clinical Psychology, 42,* 95-99.

National Institute of Statistics. (2002a). *Anuarul Statistic al României*. Bucureşti: INSSE.

National Institute of Statistics. (2002b). *Recensământul populaţiei şi locuinţelor*. Bucureşti: INSSE.

Oakland. T. (1995). 44 country survey shows international test use patterns. *Psychology International, 6(1)*, Winter, 7.

Prieler, J. A., & Raven, J. (10/20/02) The Measurement of Change in Groups and Individuals, with Particular Reference to the Value of Gain Scores: A New IRT-Based Methodology for the Assessment of Treatment Effects and Utilizing Gain Scores. *WebPsychEmpiricis*t Retrieved 10/20/02 from: http://www.wpe.info/papers_table.html

Rasch, G. (1960/1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Illinois Press.

Raven, J. (1981). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.1*: *The 1979 British Standardisation of the Standard Progressive Matrices and Mill Hill Vocabulary Scales, Together With Comparative Data From Earlier Studies in the UK, US, Canada, Germany and Ireland*. San Antonio, TX: Harcourt Assessment.

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1-48.

Raven, J. (2000a). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.3 (Second Edition): A Compendium of International and North American Normative and Validity Studies Together with a Review of the Use of the RPM in Neuropsychological Assessment by Court, Drebing, & Hughes*. San Antonio, TX: Harcourt Assessment.

Raven, J. (2000b). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1-48.

Raven, J. (2006) Lessons Learned While Developing a Romanian Version of the Mill Hill Vocabulary Test. WebPsychEmpiricist. http://www.wpe.info/papers_table.html

Raven, J., Prieler, J., & Benesch. M. (2005, in preparation). A Cross-Validation of the Item-Analysis of the Standard Progressive Matrices *Plus*, Together with a comparison of the results of Applying Three Variants of Item Response Theory. WPE

Raven, J., Raven, J. C., & Court, J. H. (1998, updated 2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1*: *General Overview*. San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3*: *The Standard Progressive Matrices*. San Antonio, TX: Harcourt Assessment.

Raven, J. C. (1939). The RECI series of perceptual tests: An experimental survey. *British Journal of Medical Psychology, XVIII*, Part 1, 16-34.

Raven, J. C. (c.1950). *Progressive Matrices (1947): Plan and Use of the Scale with The Report of An Experimental Survey Carried Out by G. A. Foulds.* London: H. K. Lewis.

Sandu, D. (1996). *Sociologia tranziţiei. Valori şi tipuri sociale în România*. Bucureşti: Staff.

Sandu, D. (1999). *Spaţiul social al tranziţiei*. Iaşi: Polirom.

Vodegel-Matzen, L. B. L. (1994). *Performance on Raven's Progressive Matrices*. Ph.D. Thesis, University of Amsterdam.

Wright, B.D. (1980) in a Foreword to Rasch, G. (1980). *Probabalistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.

Chapter 6

Lessons Learned while Developing a Romanian Version of the *Mill Hill Vocabulary* Test*

John Raven**

Abstract

Whereas Raven's *Progressive Matrices* tests have repeatedly been shown to have impeccable test properties in many different cultures, it has proved remarkably difficult to develop a range of *Mill Hill Vocabulary* tests for cross-cultural use. This has more implications for the use of tests on a cross-cultural basis than might at first sight appear since, on the face of it, nothing could be simpler than generating equivalent sets of words for use in different cultures. The present paper shows that even generating parallel versions of the same IRT-based test, in open-ended and multiple-choice formats, is fraught with difficulties. For example, the introduction of apparently acceptable distractors into multiple-choice versions of items which function effectively in open-ended format can destroy them. The relative merits of alternative computer programs for carrying out the requisite analyses are assessed, and most found wanting.

* A version of this chapter is in electronic form on the Web Psych Empiricist http://www.wpe.info/papers_table.html

** The data on which this paper is based were collected by Anca Dobrean and Camelia Rusu with the assistance of Mircea Comşa, Robert Balazsi and numerous students. The questions the study sought to answer were raised by Camelia Rusu. The analyses were carried out by Joerg Prieler and Jean Raven.

Introduction

Raven's *Mill Hill Vocabulary* (MHV) tests (of which there are several versions derived from one basic version) were developed in 1938/39 to measure Spearman's "*reproductive*" ability alongside Raven's *Progressive Matrices* (RPM) tests - which measure Spearman's *eductive* ability.

The development of the scale is described in some detail in the relevant section of the *Manual* (Raven, J., Raven, J. C., & Court, 1998). As with the RPM, a graphical version of *Item Response Theory* was used to check whether the words scaled properly, and in the same way, in each of the versions of the test (specifically when in open-ended and multiple-choice format) and, in particular, whether information contained within the distractors in the multiple-choice versions interfered with the rank order of difficulty of particular words. The causes of any variation in the shapes of the *Item Characteristic Curves* (ICCs) across versions were investigated and the items either replaced or corrected. As with the RPM, these ICCs were plotted separately for children from different socio-economic backgrounds.

When revising the sequence of words, and later some of the words themselves, in response to changing word usage and cross-cultural (particularly UK - US) differences in word usage in the 1970s, use was made of information on both the average difficulty of the words in different cultures (specifically, Australia, the US, and the UK) and in the shapes of the item characteristic curves (Raven, 1981).

These questions re-surfaced in the context of the development of the Romanian version of the test.

More specifically, could one, using the, in some ways, more sophisticated computer programs that had been developed in the interim show that the words really had the same order of difficulty when presented in multiple-choice compared with open-ended format; could one demonstrate that the words which had been developed for the "parallel" versions of the test really had the same difficulty as, and functioned in the same way as, those they were thought to parallel; and could one demonstrate that, where words were of very different difficulty in the open-ended and multiple-choice forms of the test, this was due to the presence of certain distractors.

Due to the difficulties of replicating Raven's original procedures using the methods developed by Fischer (see Raven, Prieler, & Benesch, 2005) and even generating sets of 84 3-pl ICCs of the form developed by Benesch and Prieler (see above paper) using a DOS version of BILOG

(which had been shown to approximate the Raven/Fischer curves) an attempt was made to use an alternative program - RUMM - which had been promoted as a solution to most of our prayers.

Unfortunately, (i) collection of the relevant data from a nationally representative sample of Romanians was delayed due to technical reasons and (ii) we were unwilling to invest in a full version of RUMM without an assurance that it would answer all our questions. Accordingly we used the demonstration version of RUMM, which is limited to 99 respondents and 16 items.

Data drawn from a subsample of Romanian respondents for 16 items were therefore assembled for the analyses to be reported here.

At this point it is necessary to say a little more about the MHV itself.

There are two Forms of the *Senior* version of the test, which will be our concern here.

Each of these consists of two Sets of 34, hopefully parallel, words, known as Set A and Set B.

In Form I, Set A words are presented in open-ended (OE) format and Set B in multiple-choice (MC) format.

In Form II, Set A words are in multiple-choice format and Set B in open-ended format.

To check whether either of the sets of words are parallel when presented in the same or different formats two samples of individuals are required, one of whom has taken Form I and the other Form II.

Question 1: *Can RUMM generate 33 sets of ICCs yielding as much information as the sets of 3-pl ICCs shown in Figure 12 of Raven, Prieler, and Benesch?*

Answer: "No". Figures 6.1 and 6.2 show that the only ICCs we found ways of plotting using RUMM were the equivalent of 1-pl ICCs and give no indication of variation in slope (i.e. the discriminatory power of the items) or the effects of distractors (often misleadingly subsumed under the words "guessing" or "chance") on the shapes of those curves. Thus it is impossible from them to derive any insights into item functioning or how to correct malfunctions.

Question 2: *Do the same items have similar difficulty when presented in open-ended and multiple-choice format?*

Answer: "No". Using classical (i.e. not IRT-based) indices of item difficulty, it is clear from Table 6.1 that items A20, A22, A26, B22,

Figure 6.1. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
RUMM Item Characteristic Curves
Set A, Multiple-Choice

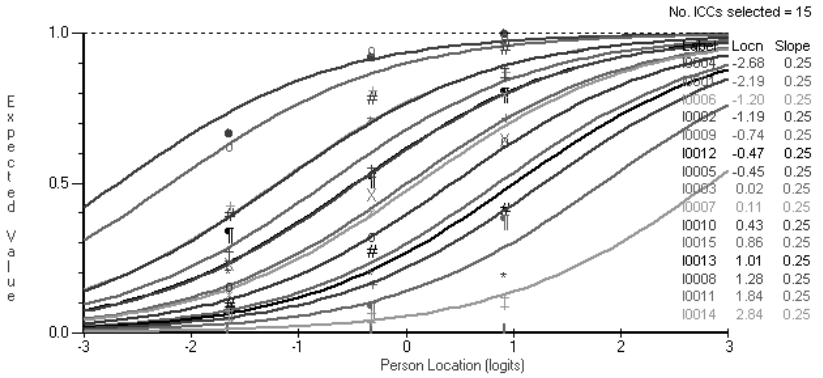
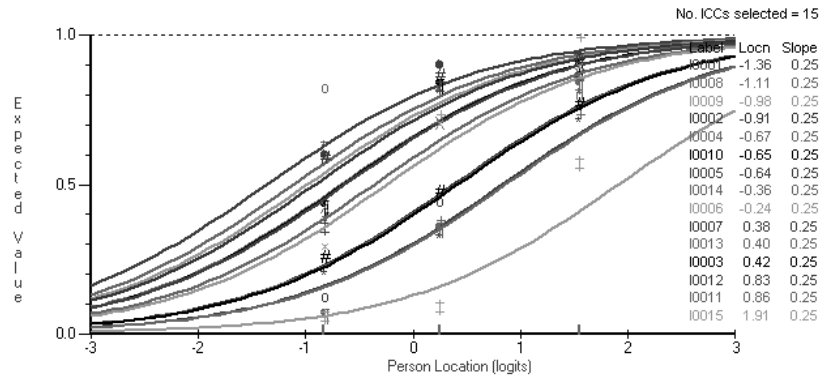


Figure 6.2. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
RUMM Item Characteristic Curves
Set A Open-Ended



B23, and B24 are very much more difficult in multiple-choice format. Something is distracting those who really know the answer!

Out of the sub-set of items included in this study, only B18 is easier in multiple multiple-choice format, presumably because of information contributed by the distractors.

Table 6.1 also shows the words that are not of equivalent difficulty in the parallel Set in the same format.

Table 6.2 shows that some words (e.g. A16, A26 and B23) have very different discrimination indices (i.e. item-total test correlations) in open-ended and multiple-choice formats. Some, e.g. A20, have poor discrimination in both formats.

Table 6.1. *Mill Hill Vocabulary Scale: Romanian Version*
Items 13-28 (Preliminary Data)
Comparative Item Difficulties of the Same Word in Multiple-Choice and Open-Ended Format and of "Parallel" Word in Other Set

	Item Difficulties (% Correct)			
	Set A		Set B	
	MC	OE	MC	OE
Item	Sample 2	Sample 1	Sample 1	Sample 2
13	83	82	59	76
14	68	75	85	60
15	41	48	44	23
16	86	72	44	42
17	52	68	35	43
18	68	61	57	29
19	40	48	42	31
20	16	79	84	76
21	56	75	65	57
22	33	70	34	73
23	13	39	16	65
24	51	38	34	60
25	23	48	83	87
26	4	64	75	64
27	27	22	32	11
28	30	17	23	20
n	92	93	93	92

Key: Difficulty of OE word very different from MC.
Difficulty of Set A word very different from Set B equivalent in same format.

Question 3: *Do the Distractor Characteristic Curves help us to find out whether some distractors confuse people who know the answer?*

Answer: "Yes". But we should first look at Figure 6.3, which presents the item distractor curves results for a reasonably well-functioning item - A15. It will be seen that choice of the correct answer increases with total score while choice of distractor 4 falls away. Thus the item works as it should, although some distractors attract no one and thus have no function.

Figure 6.4 shows that Item A17 behaves even better.

Figure 6.5, relating to item A20, shows something different. Choice of option 3 (the correct answer) falls with increasing total score, as does

Table 6.2. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
Comparative Item Discriminative Power of the Same Word in Multiple-Choice and Open-Ended Format and of "Parallel" Word in Other Set

	Item-Total Correlations			
	Set A		Set B	
	MC	OE	MC	OE
	Sample 2	Sample 1	Sample 1	Sample 2
Item	r_{it}	r_{it}	r_{it}	r_{it}
13	.40	.38	.50	.62
14	.50	.35	.51	.56
15	.31	.47	.51	.38
16	.46	.26	.60	.43
17	.51	.59	.52	.53
18	.43	.56	.67	.48
19	.56	.62	.56	.61
20	.13	.17	.62	.59
21	.64	.46	.38	.48
22	.45	.47	.42	.47
23	.42	.62	.20	.41
24	.40	.46	.26	.34
25	.35	.34	.31	.47
26	.10	.47	.55	.49
27	.58	.53	.39	.42
28	.42	.41	.32	.28
n	92	93	93	92
Cronbach Alpha	0.699	0.736	0.749	0767

choice of distractor 1, while choice of distractor 4 (a wrong answer) increases with total score.

Figure 6.6, relating to item A26, shows that distractor 1 attracts almost everyone and deflects them from the correct answer.

Figure 6.7, relating to item B22, again shows a well functioning item.

Figure 6.8, relating to item B23, again reveals how choice of a wrong answer can increase dramatically with total score.

Figure 6.3. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
RUMM Item Distractor Characteristic Curves
Item A15

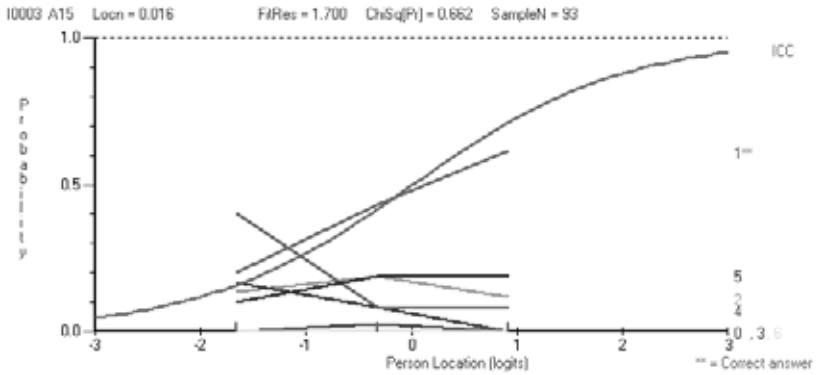
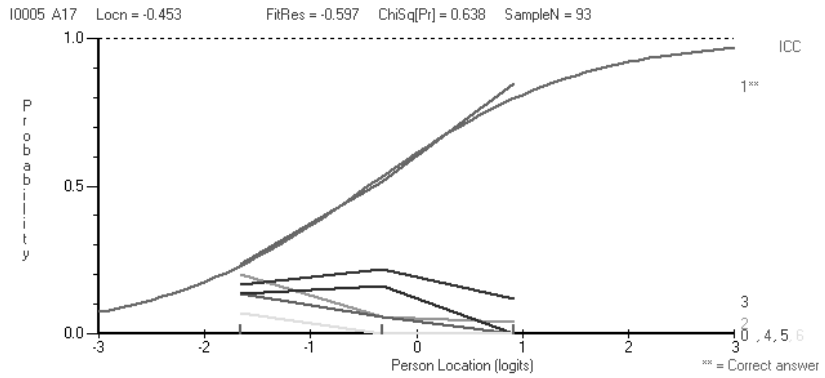


Figure 6.4. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
RUMM Item Distractor Characteristic Curves
Item A17



What do sets of 3-pl ICCs tell us?

The information assembled in Raven, Prieler, and Benesch (2005) shows that the original (1935) Raven ICCs reveal a great deal more about item functioning than do modern 1-pl ICCs … which reveal almost nothing. Nevertheless that paper shows that 3-pl ICCs do a good job of approximating the Raven curves. Although most of the differences between the 1- and 3-pl curves in that paper are not striking, it must be

Figure 6.5. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
RUMM Item Distractor Characteristic Curves
Item A20

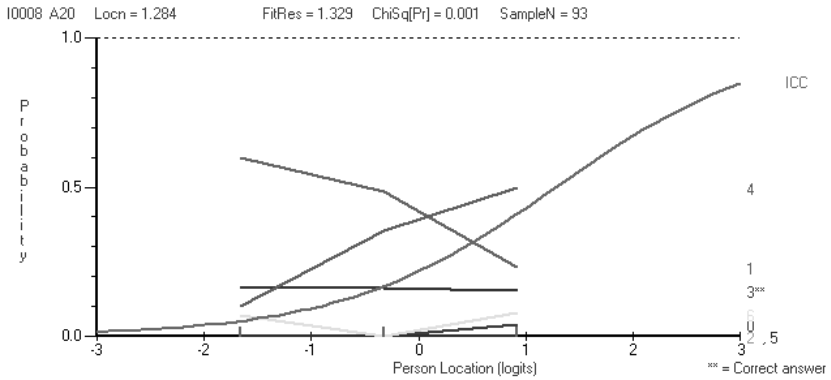
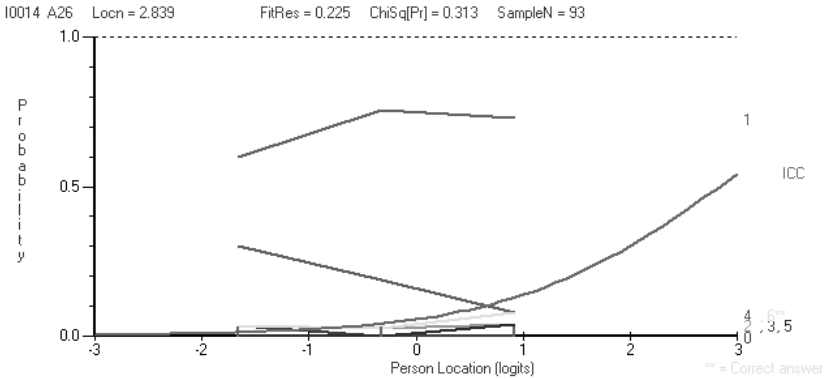


Figure 6.6. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
RUMM Item Distractor Characteristic Curves
Item A26



noted that *all* the data relate to good items which had been extensively worked over. This is why few serious defects can be discerned in the plot of 60 3-pl SPM*Plus* ICCs. The data presented in Figure 6.9 for the preliminary Romanian data for Set A in multiple-choice format (the RUMM 1-pl ICCs for which were discussed earlier) tell a different story, although it must immediately be reiterated that this is purely a methodological study based on a small sub-sample of people and items. It

Figure 6.7. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
RUMM Item Distractor Characteristic Curve
Item B22

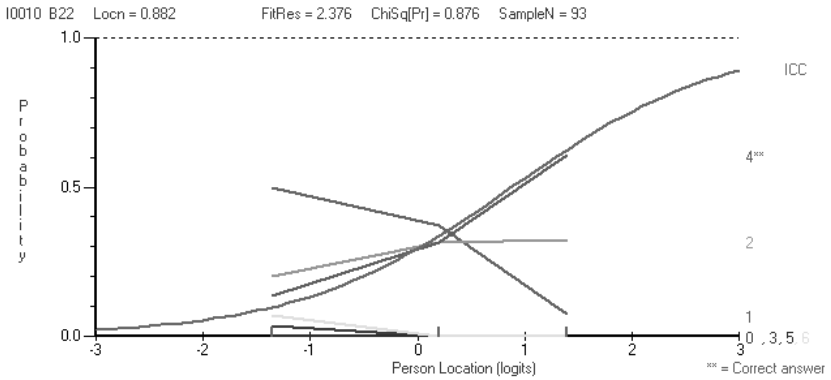


Figure 6.8. *Mill Hill Vocabulary Scale*: *Romanian Version*
Items 13-28 (Preliminary Data)
RUMM Item Distractor Characteristic Curve
Item B23

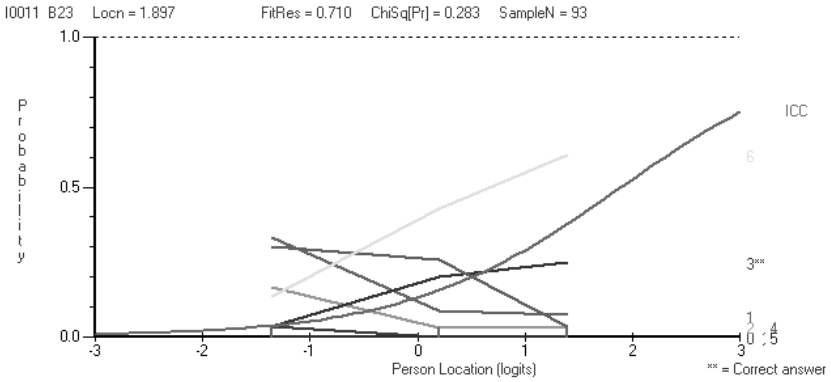
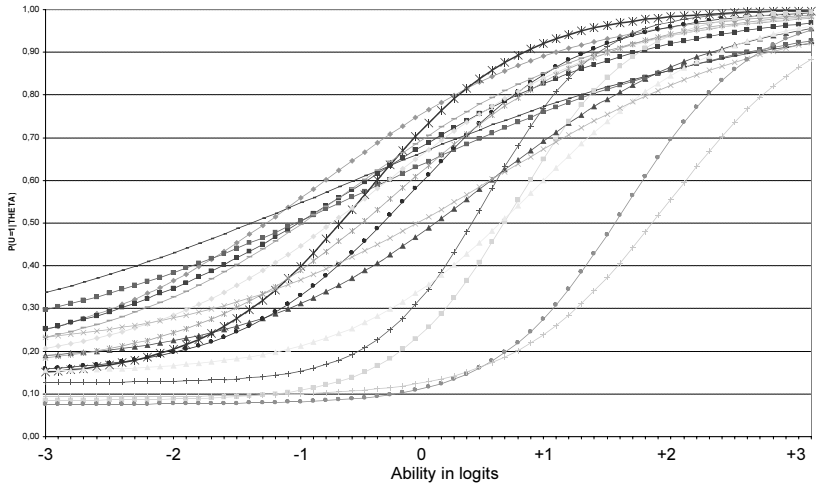


Figure 6.9. *Mill Hill Vocabulary Scale*: *Romanian Version*: *Preliminary Data*
Set A, Multiple-Choice
Items 13-28; Sub Set of Respondents
3-pl Item Characteristic Curves



is therefore extremely unlikely that the results displayed in Figure 6.9 will replicate when the full data set becomes available. Nevertheless, from a methodological point of view, the results are striking.

Although it is not possible in that Figure to identify which curve belongs to which item, it is obvious from their ICCs that some of the items are functioning very poorly. Their ICCs cross those for *all* the other items. Far too many low ability people get these items right and far too many high ability people never get them right. In other words, there is something about these items which leads low ability people to select the correct answer and something which prevents the more able from choosing it. As we have seen, plotting the *Item Distracter Characteristic Curves* enables one to become even clearer about what, exactly, was the problem with the items.

So, can the sub-tests be considered "parallel"?

It the course of discussions of the implications of the data presented above, it was suggested that the tests might nevertheless be considered "parallel" if the graphs of their *Test Information Functions* (TIF) were similar.

Test Information Function curves plot the quality of the diagnostic information provided by differences between test scores at different levels

of ability. Thus, if, as is commonly the case, the TIF curve is roughly Gaussian (often described as "normal"), it means that the test discriminates well among those of moderate ability but does a poor job among those with high or low ability. Thus, if one of the uses the data are to be put to is to differentiate among those who have been referred as potential candidates for Special or Gifted education, this is not exactly desirable. Thus, contrary to what might be expected, the ideal shape for a test information function curve might be rectangular or even bimodal (see Hambleton et al., 1991 for a fuller discussion.)

The *Test Characteristic Curves* (TCCs) for the 4 variants of the subset of MHV items and respondents discussed here are shown in Figure 6.10 and the *Test Information Function* curves in Figure 6.11.

The *Test Characteristic Curves* are not dramatically different, although it can hardly be said that they are "the same".

The same cannot be said for the *Test Information Function* curves. So it would seem that the deficits in the tests and the differences between them do show up here. It is therefore just possible that a "simple" comparison of the TIFs for different tests would enable one to decide whether they are to be considered interchangeable or not … but it would be of little help in deciding what to do about any differences that might be revealed.

****

Concluding Cautionary Note: *The above conclusions are entirely tentative*: *they are based on an analysis of a sub-set of items using a minute sub-set of the data that will become available. Absolutely no substantive conclusions should be drawn about the quality of the Romanian MHV which will eventually be published.*

Nevertheless, they heavily underline the importance of the methodological questions that were raised and indicate the analyses that would be required to answer them using the data from the full sample of respondents and items.

Figure 6.10. *Mill Hill Vocabulary Scale*: *Romanian Version*: *Preliminary Data*
Test Characteristic Curves
Items 13-28; Subset of Respondents
Set A, Multiple-Choice
Set A, Open-Ended
Set B, Multiple-Choice
Set B: Open-Ended

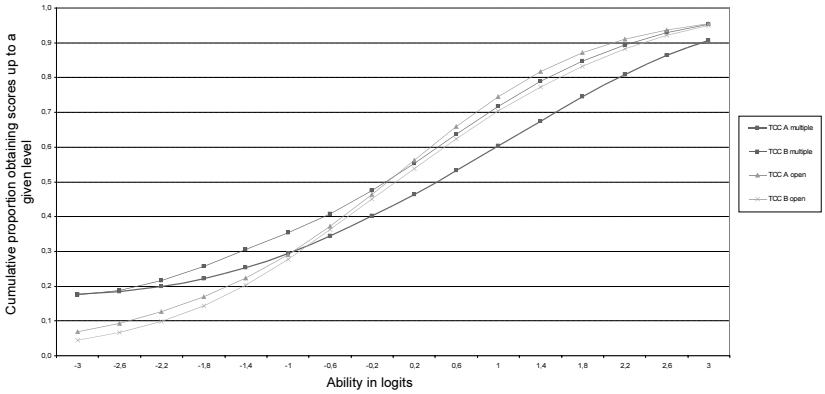
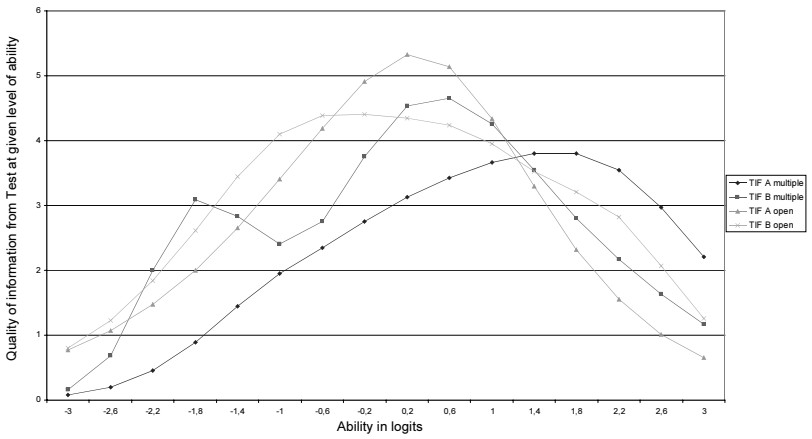


Figure 6.11. *Mill Hill Vocabulary Scale*: *Romanian Version*: *Preliminary Data*
Test Information Function Curves
Items 13-28; subset of respondents
Set A, Multiple-Choice
Set A, Open-Ended
Set B, Multiple-Choice
Set B: Open-Ended



References

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press.

Raven, J. (1981). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.1*: *The 1979 British Standardisation of the Standard Progressive Matrices and Mill Hill Vocabulary Scales, Together With Comparative Data From Earlier Studies in the UK, US, Canada, Germany and Ireland*. San Antonio, TX: Harcourt Assessment.

Raven, J., Prieler, J, & Benesch. M. (2005). A replication and extension of the item-analysis of the Standard Progressive Matrices *Plus*, together with a comparison of the results of applying three variants of Item Response Theory. http://wpe.info/papers_table.html Updated in the previous chapter of this book.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 5*: *The Mill Hill Vocabulary Scale*. San Antonio, TX: Harcourt Assessment.

Chapter 7

Problems in the Measurement of Change (with Particular Reference to Individual Change [Gain] Scores) and their Potential Solution Using IRT*

Jörg A. Prieler and John Raven

Summary and Overview

Part I of this paper reviews problems in the measurement of change and, in particular, in the calculation of change scores (such as before vs. after difference scores purporting to index variance in people's responsiveness to such things as instruction, stress, therapy and drugs).

Part II describes a new approach to the differential measurement of change among respondents having different levels of ability and, in particular, the calculation and use of change scores.

******

Some of the problems involved in assessing change, and especially differential change among high and low scoring respondents, are well known, if widely neglected in practice. These include problems arising from ceiling effects and from uneven increases in the difficulty of the items at different points in a scale.

There is, however, a much less widely appreciated, but still more serious, problem. In order to indicate the nature of that problem in this overview, it is necessary here to use technical terms loosely and in such a way that their meaning is suggested by their context. A more

* An earlier version of this paper was published in *Psiholoska obzorja/ Horizons of Psychology, (2002), 11(3)*, 119-150.
That version has for some time also been available in electronic form on the *Web Psych Empiricis*t http://www.wpe.info/papers_table.html

technically accurate statement of the problem will be found in Endnote 1. The problem is that the raw score differences that correspond to equal differences in latent ability vary markedly with the absolute difficulty of the test employed, the shape of its test characteristic curve, and the sector of that curve on which the change is measured. This is true even on tests which satisfy the requirements of the most popular versions of Item Response Theory (often loosely referred to as "the Rasch model" [see accompanying box for a non-technical explanation of this and related terms]). It is therefore impossible to draw valid conclusions about such things as the relative gains of high and low ability pupils in response to educational practice using the procedures that have been most widely employed. It also follows that "before vs. after" difference scores designed to assess *individual* responsiveness (such "learning potential" [i.e. ability to learn from instruction] "sensitivity to noise" or "responsiveness to drugs of type A") signify different things at different points in the distribution.

Part II of this paper outlines a methodology to overcome this problem. Although grounded in Rasch theory, it is widely applicable to tests which are not Rasch homogeneous. It is based on two ingenious observations. The first is that the same item administered at two points in time must constitute a homogeneous, if miniature, Rasch scale. The second is that changes in item parameters on items which constitute a Rasch scale can be used to index changes in *people.*

The result is an extremely flexible, and widely applicable, set of procedures for assessing change.

Part I: Problems in the Measurement of Change

Introduction

The main focus of this article is on the closely related problems of (a) the measurement of *differential* change in groups – for example among people of different levels of ability in response to some treatment (e.g., did high-ability students gain more from an educational enrichment program than low-ability students?), and (b) the calculation and interpretation of *individual* "change" scores (for example, in the measurement of "Learning Potential" or "the ability to learn", by subtracting the individual scores people achieved on Raven's *Progressive Matrices* before and after a period of training). However, in order to more fully illustrate the problems which provoke a need for this discussion and new methods for

overcoming them, some more general problems arising in attempts to assess change will first be discussed.

The problems involved in measuring change in psychological characteristics will, in the remainder of this article, be discussed under the following headings (see also Endnote 2):

1. Problems arising from floor and ceiling effects.
2. Problems arising from the frequently encountered need to use a different and more difficult test to assess performance after an intervention, such as an educational program, because the knowledge and ability of *all* concerned has improved dramatically as a result of the intervention.
3. Problems arising from the available tests not yielding equal-interval scales.

   There are two sets of problems here:

   a. Problems arising from uneven probit distributions within tests constructed using classical test theory.
   b. Problems arising from the fact that equal raw score differences among low and high ability individuals do not imply equal differences in latent ability. This applies even to tests conforming to most popular versions of Item Response Theory (IRT) and Rasch scaling [See accompanying box for an approximate, non-technical, explanation of these terms]. Because the implications of this are both unexpected and important, and because it is these problems which the methodology discussed later makes it possible to solve, the bulk of this paper will be devoted to highlighting this particular problem.
4. Problems arising from a preoccupation with single-variable assessments which themselves stem from the obvious problems involved in employing a wide array of classical multivariate scales, that is to say, problems arising from the use of the insufficiently *comprehensive* assessments which a preoccupation with classical measurement scales entails.
5. Problems arising from the low reliability, construct validity, and predictive validity – and thus meaningfulness – of *differences* between *individual* scores before and after some treatment (such as training or stress) – that is to say problems having to do with the meaningfulness of *individual* "gain" or "loss" scores as indices of some deeper personal characteristic – i.e. as measures of such

In the main text of this article we have tried to limit ourselves to the use of terms that we believe are becoming generally familiar to psychologists. However, to assist those who have not yet acquired such a nodding acquaintance, we will try to indicate what the terms mean in endnotes.

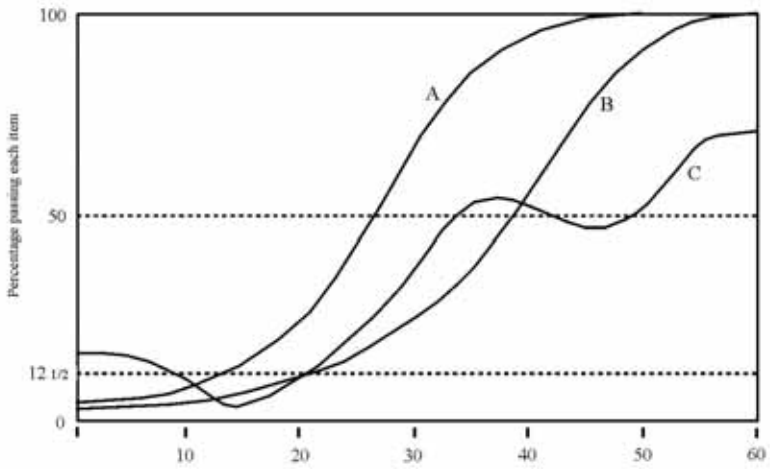
In the paragraph to which this note refers, "classical test theory" refers to the methodology most widely employed to construct what are deemed to be unidimensional tests. This usually involves intercorrelating the items and subjecting the resulting correlation matrix to some form of factor analysis.

The term "Item Response Theory" (IRT) refers to attempts to develop tests in which the items form a sequence in which respondents endorse, or answer correctly, all items up to a certain point and then reject, or get wrong, all subsequent items. Examples include Guttman scales in the "attitude" domain and the Raven Progressive Matrices and the British Ability Scales in the "ability" domain. In the physical domain, perfect scales of this type include the use of meter sticks to measure length: everything that is 15cm. tall is more than 14cm, 13cm, 12cm etc. And less than 16, 17, etc. Note that it would not make sense to seek to establish the unidimensionality of a meter stick by intercorrelating and factoring the items.

An appropriate methodology for constructing equivalent scales in the psychological domain began to emerge in the UK in the mid 1930s (and was used in the development of the Raven Progressive Matrices in 1935), was given mathematical form by Rasch in Denmark in the late 1940s (Rasch, 1947), and was popularised by Wright (1968) and others (e.g., Lord & Novick, 1968) in the US in the 1960s. Theoretical work in the area grew apace, e.g., Birnbaum (1968); Fischer (1974); Embretson (1999).

The most fundamental requirement of IRT is that test constructors somehow demonstrate that the graphs of the way in which the proportion of respondents getting each item right varies with ability are systematic in themselves and display a systematic relationship to the graphs for other items. In Figure 7.1, graphs A and B display this relationship while graph C does not. (See Raven, Raven, & Court 1998a or Hambleton, Swaminathan, & Rogers, 1991 for a fuller discussion of this Figure).

Figure 7.1. **The Hypothetical Behaviour of Test Items**



(Reproduced from Raven, J., Raven, J. C., & Court, 1998a)

In more formal treatments "ability" is indexed, not by the total score on the test, but by "latent trait" score. A "latent trait" is loosely equivalent to "underlying factor" in classical test theory.
A more technically correct treatment of these issues will be found in Endnote 3.

things as "learning potential", "sensitivity to stress", or "strength of reaction to a drug".

1. Problems arising from ceiling effects

Simple "Ceiling Effects" arising from the inability of respondents to demonstrate their prowess because there are either insufficient difficult items in a test to allow them to make their capability known or insufficient time for them to demonstrate it are well known. Yet the mistaken conclusions which the use of tests with too low ceilings (or too high floors) have induced researchers who have sought to document change or compare the relative merits of alternative treatments (such as different types of educational program) to draw are pervasive and generally pass unnoticed. We will begin with simple examples (which might be considered trite were they not so common). Later, we will illustrate some of the effects which are more difficult to detect. Note that it is their detection which poses the problem. For, despite the recommendation – couched in the strongest possible terms – from the APA Task Force on Statistical Inference (APA, 1999) that researchers should carefully examine – indeed graph – their raw data before subjecting it to statistical treatment, many researchers fail to do so. Yet, once data have been summarized using widely accepted methods, it is extremely difficult for even critical and motivated readers to detect many of the "obvious" ceiling effects we will describe.

Example 1.

We begin with a simple, but all too common, example, which we will return to and elaborate later.

Suppose two groups of managers have been identified – an "average" group and a group of "superstars". Both attend a seminar designed to enhance their capability. One hypothesis might be that the average managers will *gain* more than the superstars because the latter will already know what is being taught. To test this hypothesis we might construct a test of managerial knowledge. We arrange for all participants to be tested before and after the seminar. We then, like many other researchers, find that the average gain by the less able managers is greater than that among the more able. Our hypothesis is confirmed.

But what might actually have happened? Suppose the superstars already knew the correct answers to nearly all the questions posed at the pretest, but had indeed gained a great deal from the seminar. They would

then not be able to demonstrate the gains they had made. It might be supposed that this problem could be easily solved by lengthening the test. But that is not the case. For a start, if we simply added numerous items suited to the superstars we could easily reverse our conclusion and demonstrate that they had gained a great deal "more" than the average managers.

This apparently simple problem turns out to be much more complex than meets the eye. Failure to address it means that the interpretation of a great deal of research (such as that designed to find out whether more or less able students gain more from educational enrichment programs or from differences in the organization of education, such as "streaming vs. mixed-ability teaching") is seriously misleading. Thus, one of the objectives of this paper is to unpack the background to this problem more fully and then show how it can be solved.

But now let us suppose something else. Let us suppose that the benefits the superstars gained from coming to the seminar were of a different qualitative nature to the benefits derived by the average managers. Suppose, for example, that the main benefits the superstars got from the seminar came from their interactions with other superstars and not from the lectures and case studies. They could then not be expected to show up on any unidimensional test of managerial knowledge. Yet the attempt to use a collection of unrelated items to trawl for possible effects of the seminar would pose enormous problems for traditional forms of data reduction and significance testing. We will return to this problem under heading (4) below because failure to focus on *comprehensiveness* in assessment is perhaps the most important problem currently encountered in psychometrics … and it is one that the methodology to be discussed later helps us to overcome.

Example 2.

We turn now to another example of the methodological problems posed by the ceiling effect … an example which will, in the end, help us to illustrate the sources of the measurement problem we aim to highlight and how that problem is to be overcome.

As is now becoming well known, largely as a result of the publicity given to it by Flynn (1984, 1987, 1999), the scores of random samples of the population on most multicomponent measures of "general intelligence" have been going up fairly dramatically over time. To give a general indication of the rate and magnitude of the increase, Flynn cites a figure of one standard deviation per generation. The effect, calculated

in SDs per generation, is greatest on measures of reasoning – or, more correctly, "*eductive"* ability – and least on measures of knowledge or routine skills – referred to as *reproductive* abilities.

The increase in *eductive* ability scores, as measured by Raven's *Standard Progressive Matrices*, is shown in Figure 7.2. The Figure graphs the percentile norms obtained by adults of different ages (and thus dates of birth) on the *Standard Progressive Matrices* (Classic Form) from one sample of the British population tested circa 1942 and another tested in 1992. The approximate age of people born in different years in the two samples is shown below the graphs.

It is immediately obvious that the raw scores of the "less able" – the scores of those at the 5th and 10th percentiles – have gone up more than the raw scores of the more able – those scoring at the 90th and 95th percentiles. However, it is equally obvious from data presented in this form that the failure of the scores of the more able to increase more is, at least in part, a product of the test ceiling which, with 60 items, does not allow the more able of those born more recently to reveal what they can do. But the point to be made here is that numerous researchers, looking only at summarizing statistics without investigating the possibility of a ceiling effect, concluded that the scores of the less able have been increasing faster than those of the more able.

When data from a more difficult version of test – the *Advanced Progressive Matrices* (APM) – are examined it is, as shown in Table 7.1, immediately obvious that the scores attained by the more able *have* also increased dramatically. Unfortunately, we still do not know whether the effect is greater or less than that among the less able because a more difficult test with different test characteristics has been used. Furthermore, it would appear that it is not possible to draw even tentative conclusions from the APM data because the increase has been so great that it has run into the ceiling of even on *that* test (which has only 36 items).

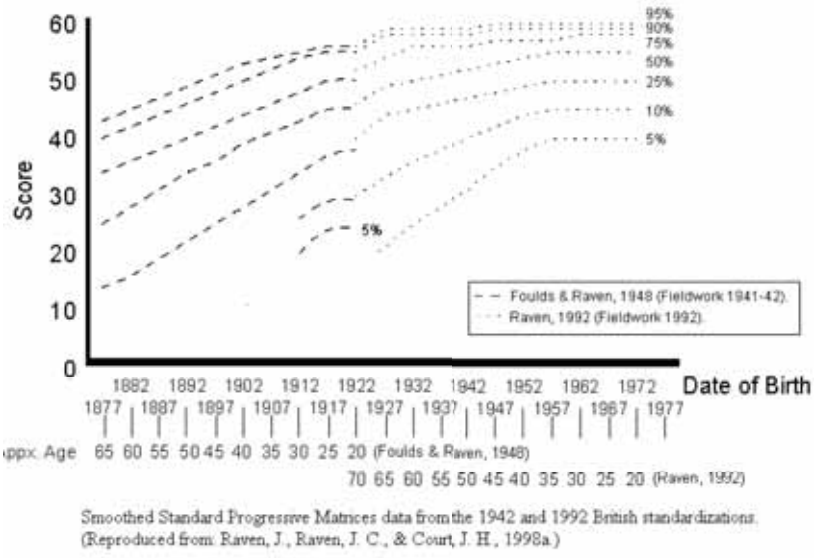
We will return to this problem later. But here it is important to note that, in an attempt to answer this question, Teasdale & Owen (1989) examined the latent trait scores which people obtained on a test which required respondents to complete geometrical shapes. Although, in the abstract to their paper, they state that "we find no evidence of gains at the higher levels", it is truer to the general tenor of their findings to say that they concluded that, although there were gains among the more able, the gains were greater among the less able.

The study is of particular importance in the context of this article because (i) the analysis was conducted using the latent trait scores

emerging from a Rasch analysis and (ii) particular care was taken to check for the possibility of a ceiling effect.

Before examining the possibility that a concealed ceiling effect may nevertheless have been operating, it is necessary to compare the

Figure 7.2. Standard Progressive Matrices
100 years of eductive ability in Great Britain



Smoothed Standard Progressive Matrices data from the 1942 and 1992 British standardizations.
(Reproduced from: Raven, J., Raven, J. C., & Court, J. H., 1998a.)

Table 7.1. **Advanced Progressive Matrices, Set II**
Comparison of 1992 and 1962 UK adult percentile norms.

	Age in years					
	20		30		40	
Percentile	1962	**1992**	1962	**1992**	1962	**1992**
95	24	**33**	23	**33**	21	**32**
90	21	**31**	20	**31**	17	**30**
75	14	**27**	12	**27**	9	**26**
50	**9**	**22**	**7**	**22**	**--**	**20**

Note. The 1962 data (previously published in Raven, J. C., 1965) were estimated from the work of Foulds & Forbes, which was also published in Raven, J. C. (1965).

Since the test has 36 items and 8 options per item, scores of 6 or less verge on the chance level. There was therefore no point in publishing the lower percentiles for 1962.

The 1992 data come from Raven, J., Raven, J. C., & Court, J. H. (1998b).
Reproduced from Raven (2000b).

construct validity of the test used by Teasdale & Owen with that of Raven's *Progressive Matrices* lest differences in the time trends on the two tests arise from this source. This is important because Raven (2000b) concluded from a review of the literature – which included the studies of Thorndike (1975), Schaie & Willis (1986), and Flynn (1999) – that scores on tests measuring different components of "cognitive ability" have not all increased to the same extent. Scores on measures of *eductive* ability, whether measured by verbal or nonverbal tests, are increasing at about 1 standard deviation per generation, those on measures of *reproductive* ability hardly at all, and those on tests which tap both components of cognitive ability at rates which depend on their factor loadings on these two more basic abilities. This conclusion was later confirmed in Flynn's (2000) study of the subscales of the Wechsler test.

Since scores on the test used by Teasdale and Owen increased at .5 of a standard deviation per generation it seems likely that it tapped both eductive and reproductive abilities – in which case, as others have found, one would expect to get less clear-cut results.

But is there a possibility of a concealed ceiling effect? Indeed there is, for, although this was a power test, it was also timed. As a result, more able respondents may not have been able reveal what they were actually able to do.

It is, in fact, always a mistake to mix up speed and power (as in this case) when attempting to measure change. There are two reasons for this:

(1) Because of the time limit, many people fail to reach the items at the end of the test. As a result, it is impossible to calculate true item parameters for these items – including their difficulty levels.

(2) A "time limit ceiling effect" arises directly from the time needed to answer the questions. It is easiest to see this from an example. Suppose one administers a test composed of 60 very easy items with a five minute time limit. Suppose further that a very able person is just able to answer all 60 items correctly in this time. That is, he requires five seconds to answer each item and turn the page. He then attends a training program which greatly increases everyone's ability. To compensate for this, the researcher lengthens the test to 80 items. Our respondent is now much cleverer than he was, *but he still requires five seconds to answer the questions and turn the page.* His score is still 60!

2. Problems arising from the fact that it is often necessary to use a different and more difficult test after an intervention

We have already encountered this problem twice. The first was when we saw that a managerial development program might enhance the capacity of the more able beyond what the test was able to measure. The second was when we saw that the problem of documenting the relative change in the sores of the more and less able sectors of the population on the Raven's *Progressive Matrices* over time became problematic because it involved the use of tests which initially failed to discriminate adequately at the bottom end of the distribution, then provided adequate discrimination, and then failed to discriminate adequately at the top end. Although data collected with a more difficult test were available and did reveal an increase at the top end, the data could not be directly converted to an appropriate metric to answer the question of whether the increases were in some sense uniform across all ability levels.

These are but particular examples of the very general problem involved in documenting the long-term (longitudinal) effects of educational enrichment programs over time scales in which it is necessary to use tests of very different difficulty levels to test the birth cohort at different ages. (It is of more than passing interest to note that the practical problem that led Rasch to formulate a mathematical version of IRT was to identify the long-term effects of an experimental reading program when different tests had [necessarily] been administered at different ages as the pupils progressed through school [Rasch, 1947, 1960/1980].)

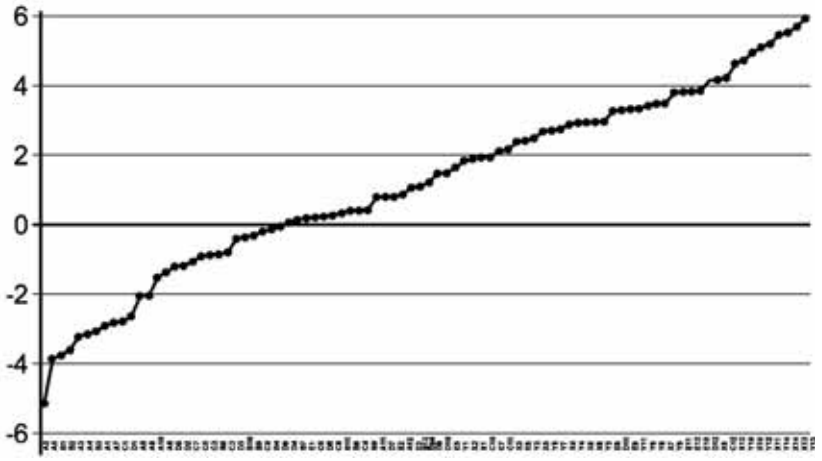
3. Problems arising from the available tests not yielding equal-interval scales

A. Problems arising from uneven probit distributions within tests constructed using classic test theory.

This problem may be illustrated by reference to Figure 7.3. This shows the item difficulties, expressed in Rasch logits, of 84 new items which were developed whilst preparing a test (the SPM ***Plus)*** to restore the discriminative power at the upper end which the *Standard Progressive Matrices* test had when it was first published, but which has, as illustrated in Figure 7.2, been eroded by the secular increase in scores.

A glance at Figure 7.3 reveals several plateaus. At these points, people's raw scores increase by 1 for each of these items that they get right. Yet, clearly, the difference between the levels of latent ability indexed by these raw scores is minimal.

Figure 7.3. *Standard Progressive Matrices* **Plus**
1996 Item Equating Study
Item Difficulties in Logits
60 Parallel items and 24 additional items



(Reproduced from: Raven, J., Raven, J.C., & Court, J.H. 2000.)

As Carver (1989) in particular has emphasized, this has led to some very misleading research conclusions. For example, as individual children get older and hit one of these plateau their raw score increases rapidly. This has contributed to the notion that there are leaps and plateau in intellectual development – i.e. times (or stages) when scores increase quickly and times when they do not. It follows that that the meaning of raw score differences, both between people and over time, must depend very much on the slope of the particular sector of the graph that is operative. And this is true despite the frequent complaints of researchers whose work focuses on particular age or ability groups to the effect that there are "too many" irrelevant items and "not enough" items in the range in which they are most interested.

At this point, it is instructive, in order to highlight further difficulties arising from the routine application of summarizing statistics to data sets in which variations in the shapes of the distribution are ignored, to return to Figure 7.2 and reexamine Flynn's claim that the rate of increase in SPM scores amounts to about one Standard Deviation per generation. As can be seen, the 50th percentile (which would, if the distributions were Gaussian, correspond to the mean) for people born in 1877 was 24. That for people born in 1972 was 54. So the actual increase in median

raw scores over the century covered by the graph was 30. However, expressing this in SD units presents difficulties. As is clear from the Figure, the distributions are not Gaussian and vary with date of birth. Every student of psychology knows how to calculate Standard Deviations using SPSS or other statistical package. And every student of psychology knows – or used to know – that 68% of the scores are encompassed within the range of mean ± 1 Standard Deviation. It follows that the standard deviation of any data set can be read off from the kind of data displayed in Figure 7.2. This can be done by estimating the range of scores which encompass 34% of the population scoring above or below the mean. As can be seen from the Figure, this yields an estimate of the SD for those born between 1900 and 1930 of about 12 if estimated from scores below the median and 8 if estimated from scores above the median. Thus the increase of 30 could be roughly expressed as rather less than one SD per each 30 year generation over the period covered by the data. However, if one estimates the SD from those born most recently and from the variance above the median, the estimate one obtains for the SD is only 3 – which gives the increase over the period covered by the Figure as 10 SDs. Of course, this is not the end of the story, because the test ceiling has depressed both the median score and the variance among those born more recently. Extrapolation of the curves – and, as has been mentioned, we know from our work with the APM that such extrapolation is justified – yields an estimate of the true median for those born in 1972 and tested in 1992 as 70. Thus the "true" increase in median score over the period covered by the data is 46, not 30. And the "true" SD is 10. So the "true" increase per 30-year generation over the 3.1 generations covered by the data is 1.5 SDs per generation. Or .9 if one takes a generation as being 20 years.

It follows that attempts to correct for irregularities in the distributions of raw scores by applying the data-reduction techniques routinely taught in elementary statistics classes, and demanded by most journal editors, are likely to yield very misleading results.

B. Problems arising from the fact that equal raw score differences among low and high ability individuals do not imply equal differences in latent ability.

To introduce a discussion of these problems we may return to our attempt to overcome the difficulty posed by our hypothetical sponsor's request that we document the relative gains made by average and superstar managers as a result of a management-development program.

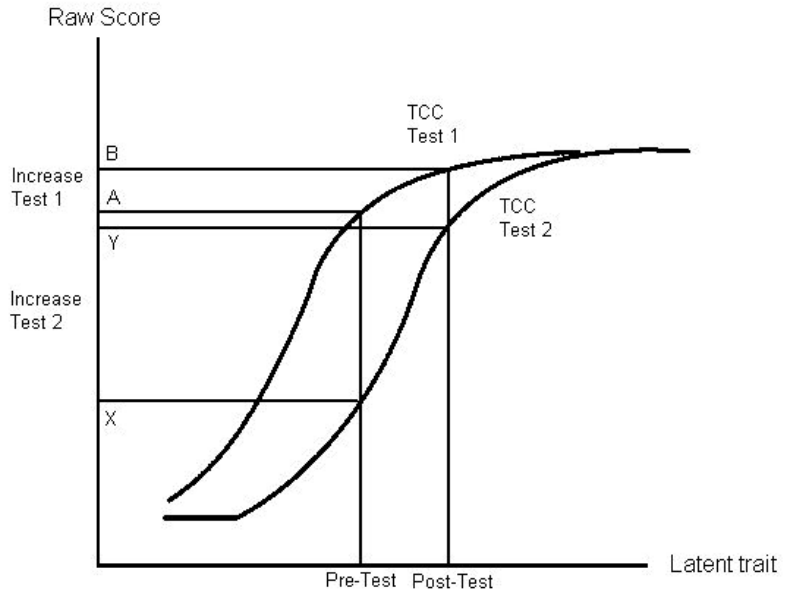
Our next step will be to elaborate on our earlier observation that, very surprisingly, the problem cannot be solved by developing a more difficult test, even a test conforming to the Rasch model.

Figure 7.4 illustrates the problem for high ability personnel and Figure 7.5 for low ability personnel. If we employ a test having the Test Characteristic Curve shown on the left in Figure 7.4, the mean scores of the high ability group increase from A at the pretest (i.e. before training) to B at posttest (i.e. after training). This is a relatively small increase. But if we use the more difficult test shown on the right, the same increase in score on the latent trait of the high ability group shows up as a *huge* increase in raw score, moving from X to Y.

As can be seen from Figure 7.5, exactly the opposite effect occurs at the other end of the scale. The apparent increase in score from pretest to posttest is huge on Test 1 and trivial on Test 2.

Putting the two cases together, it is obvious that, if the researcher employs Test 1 to assess the impact of the course, the relative gains of

Figure 7.4. *Illustration of Changes in Raw Scores on "Easy" and "Difficult" Measures of Managerial Ability for Identical Changes in Latent Ability*
High Ability Personnel Only



(Reproduced from: Raven, J., Raven, J.C., & Court, J.H. 2000.)

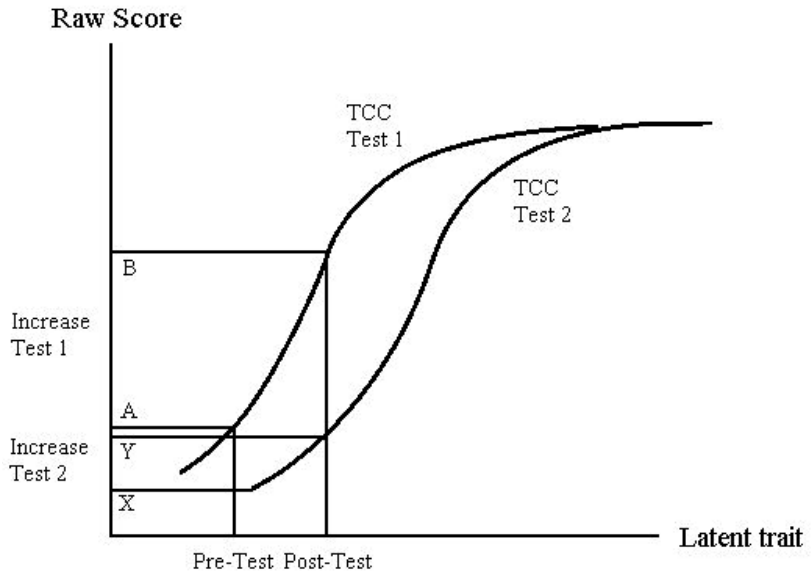
the low ability group are huge while those of the high ability group are trivial. On the other hand, if the researcher employs Test 2, exactly the opposite findings emerge.

The general, and vitally important, conclusion which emerges from these examples is that the apparent magnitude of any real increase in latent ability arising from a developmental experience or natural change over time depends (a) the general difficulty level of the test relative to the ability tested and (b) the distribution of the item parameters relative to the interval on the latent trait where change occurs.

This makes it virtually impossible, without employing the techniques to be described below, to make any meaningful statement about the *relative* magnitude of gains or losses of high, medium, and low ability groups.

More specifically, it follows from these examples and this general conclusion that we cannot solve the problem of documenting the relative gains made by high and moderate ability managers by including more difficult items and eliminating easier ones. That would have precisely the

Figure 7.5. *Illustration of Changes in Raw Scores on "Easy" and "Difficult" Measures of Managerial Ability for Identical Changes in Latent Ability*
Low Ability Personnel Only



(Reproduced from: Raven, J., Raven, J.C., & Court, J.H. 2000.)

effect of shifting the Characteristic Curve for the test being used from the curve on the left to that on the right.

But what about using a test with a *linear* Test Characteristic Curve?

Observation of the plateau in the graph in Figure 7.3 led to the elimination of some of the contributing items. This resulted in the 60 item test whose item difficulties are illustrated in Figure 7.6.

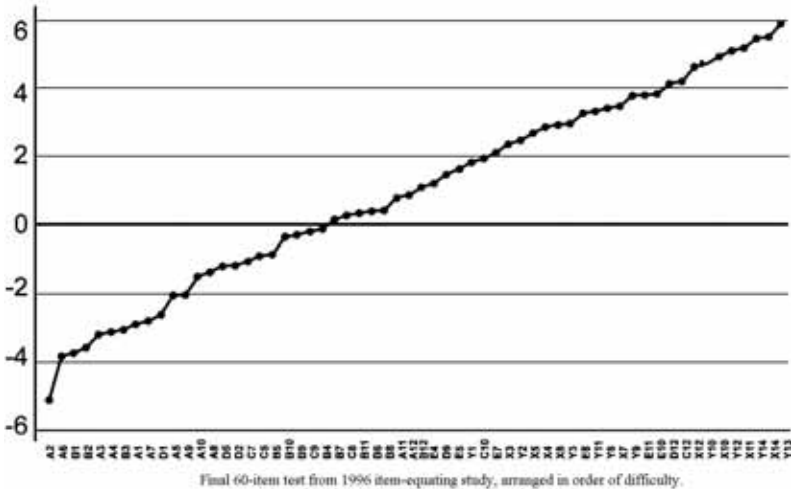
On the face of it, this provides a solution to our problem – provided such a test were used at both pretest and posttest. Unfortunately, not only does conformity to the Rasch model not ensure such an equal-interval scale, there can be no guarantee that, just because the overall distribution is as illustrated, the distributions for different ability groups will be similar.

The seriousness of this problem may be illustrated from another real data set.

The within-age distributions of the *Classic* SPM (not the SPM ***Plus,*** some of the results from the development of which were discussed earlier) are reproduced in Figure 7.7.

It is immediately obvious that these within-age distributions are bimodal and anything but Gaussian. According to a personal communication

Figure 7.6. *Standard Progressive Matrices* **Plus**
1996 Item Equating Study
Item Difficulties in Logits
Final 60 items arranged in order of difficulty



(Reproduced from: Raven, J., Raven, J.C., & Court, J.H. 2000.)

from Robert Thorndike, this is also true of the within-age within-subscale distributions on many multiple-component "intelligence" tests.

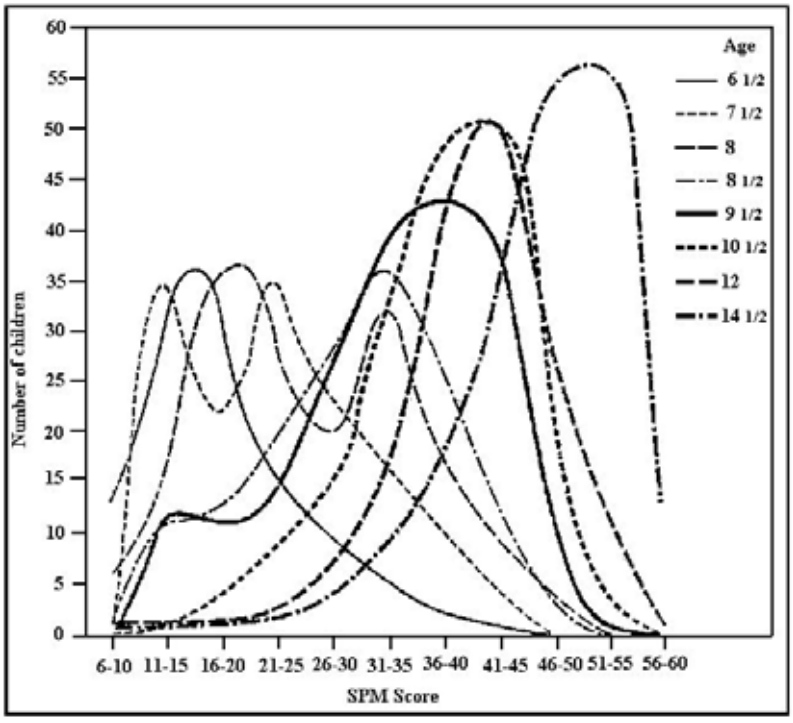
What if they were cumulated. Would we not then obtain a Gaussian distribution? The cumulated distribution is shown in Figure 7.8.

In the light of this example it is difficult to see how the overall distribution of scores on any test developed to yield the within-age Gaussian distributions that are required for all attempts to solve the problem problems posed by the differential measurement of change along the lines so far reviewed can yield an acceptable overall distribution.

******

At this point it is instructive to return to Teasdale and Owen's observations. Although, in their text, they draw attention only to the fact that the scores of high ability respondents had increased hardly at all, it

Figure 7.7. Standard Progressive Matrices *(Classic Form)*: Distribution of Raw Scores for Eight Age Groups



(Redrawn from: Raven1981)

Figure 7.8. Standard Progressive Matrices (Classic Form)
1979 British Standardisation
Overall Distribution of Raw Scores

Based on sample of 3,466 children aged 6 to 15.

(Reproduced from: Raven1989)

is obvious from Figure 7.1 in their paper that the scores of the very lowest ability groups also increased hardly at all. We do not, of course, dispute their general claim that, whereas there has been *no* increase in the raw-score equivalents of the 95th percentile, there has been considerable change in the raw score equivalents of the 5th percentile. We wish only to draw attention to the fact that *owing purely to the shape of the Test Characteristic Curve* for the test they used there actually appears to have been no increase in the scores obtained by the very least able either. (As an aside it may, however, also be observed that the Standard Deviation of the test used by Teasdale and Owen has clearly declined between the two dates for which they present data, for the TCC for the later testing is much steeper. This reduction in discriminative power, while true of the RPM more recently, is not evident in the data reported by Bouvier [1969] for many tests administered to recruits to the Belgian army from 1958 to 1967.)

4. Problems arising from a preoccupation with single-variable assessments – a preoccupation itself arising from the difficulty of measuring change using classical multi-variate scales, that is to say, problems arising from the use of insufficiently comprehensive assessments

At this point we may return, once again, to our objective of comparing the benefits our superstar managers got from attending a managerial seminar with those obtained by average managers.

Earlier, we noted that these might differ qualitatively from those obtained by the average managers and would therefore not show up on any unidimensional test of "managerial knowledge".

The use of insufficiently comprehensive evaluation packages inadequately tailored to the objectives and practices of the educational programs to be compared and their desired and desirable, and undesired and undesirable, effects on different kinds of student has resulted in a plethora of comparative evaluation studies which must be considered not only incompetent but also unethical (Raven, 1991, 2000a; Raven & Stephenson, 2001). To briefly cite one example, numerous evaluations of "open" or "progressive" educational programs have shown that they do not increase the reading, writing, or arithmetical skills of the participants as conventionally measured. But the main objectives of most programs of "open" or "progressive" education did not lie in this area: they had to do with the enhancement of self-confidence, the ability to communicate, the ability to work with others, and, above all, promoting the development of *diversity* – of different talents in different children. Furthermore the main *disbenefits* of traditional forms of education lie precisely in their destruction of positive self-images, their breeding of feelings of trained incapacity, and their creation of monocultures of mind instead of diversity. It follows that no comparative study which does not investigate such potential benefits and disbenefits can be viewed as competent or objective. Yet these inadequate evaluations – whose main failing is, above all, a lack *comprehensiveness* – have led to the closure of almost all "open" or "progressive" education programs. This not only has a seriously damaging effect on students who have the potential to develop the diverse high-level competencies which these programs might have nurtured, the programs that are denigrated are the only programs that nurture the competencies that the ex-pupils will require if they are to change our society in such a way that our species will have a chance of survival (Raven, 1994, 1995;

Raven & Stephenson, 2001). It is difficult to envisage anything that could be regarded as more unethical.

It emerges that the hallmark of scientific objectivity is the *comprehensiveness* of the assessment, not the accuracy of an assessment on a single variable. One factor contributing to the neglect of this in the past has been a preoccupation with unidimensional, multi-item, *scales.* How could one, in any practical study, aim at the kind of comprehensiveness that would stand up to the profession's demand for statistical significance testing without envisaging a huge battery of as-yet-to-be-developed tests which would, even then, not enable one to answer the question of whether most individual students had grown and developed in idiosyncratic but vitally important ways?

Nothing could better illustrate the need for two things. One is the developments to be described later. The other is a psychometric model that it is beyond the scope of this paper to discuss, but which is outlined in Raven & Stephenson (2001) and related publications. (Readers may, however, be interested to learn that Spearman noted this problem in 1926. He wrote: "Every normal man, woman, and child is … a genius at something … It remains to discover at what … This must be a most difficult matter, owing to the very fact that it occurs in only a minute proportion of all possible abilities. It certainly cannot be detected by any of the testing procedures at present in current usage. But these procedures are capable, I believe, of vast improvement".)

5. Problems arising from the low reliability, construct validity, and predictive validity – and thus meaningfulness – of differences between individual scores at different points on the test characteristic curve before and after some treatment (such as training or stress) – that is to say problems having to do with the meaningfulness of individual "gain" or "loss" scores as indices of some deeper personal characteristic – i.e. as measures of such things as "learning potential", "sensitivity to stress", or "strength of reaction to a drug".

Three sets of problems are to be reviewed under this heading: (i) those arising from the low reliability of the individual change scores; (ii) those arising from the fact that, even on tests constructed using IRT, individual change scores are highly negatively correlated with initial score, and (iii) the fact that *differences*, even when equal in terms of the construct validity of the underlying measures, may have very different meaning or

significance at different points in a scale, that is, the *gain* scores may lack uniform interpretation or construct validity.

Before moving on it may help the reader to understand the issues if we first review what is perhaps the most widely discussed attempt to utilize "gain" (individual change) scores in research expected to have major practical applications, namely that dealing with the enhancement of cognitive ability.

Many authors (e.g., Guthke, 1982; Budoff, 1973; Budoff, Corman, & Gimon 1976), but especially Feuerstein (1979) and Feuerstein, Klein, & Tannenbaum (1990), have proposed that the ability to profit from being taught how to solve problems should have higher predictive validity than straight scores on measures of "problem-solving ability". Although some researchers, perhaps most notably Guthke & Wiedl (1996), have developed more sophisticated measures, this ability, generally termed "learning potential", has typically been measured by calculating the *change* in individual respondents' RPM scores before and after a training program such as Feuerstein's "instrumental enrichment" program. In other words, people's pretest scores have typically been subtracted from their posttest scores to yield "gain" scores, and these gain scores have been presented as measures of the individual's ability to learn or "learning potential".

What we are about to discuss is the meaningfulness of such individual "gain" or "change" scores. We will later discuss ways in which the problems we will elaborate can be overcome. But first we should link what we are about to say back to the problems already discussed because it is not only the meaning and value of the individual measures of "learning potential" that needs to be examined. Researchers in the area also frequently make the claim that "low scoring" or "disadvantaged" children gain more from training than do the more able. We have already seen that it is difficult it is to substantiate such claims. Now we will show that there are still more intractable problems.

5(i). Problems arising from the low reliability of the individual change scores.

Assessing the role of error in measurement is always problematic. It becomes more problematic in change scores because it is involved twice – that is, in both the pre-test and post-test scores. Worse, these errors in both are correlated! In fact, it has long been recognized that there is an apparent paradox in this area – because, as the correlation between the pretest and posttest measures decreases, the relative error in the

change score decreases (see Lord, 1963, for the classical formula for the reliability of change scores). Worse, as Bereiter (1963) pointed out, a low pretest-posttest correlation leads to difficulties in interpreting the meaning of "change", because it indicates that the tests do not measure the same dimension! However, Embretson (1991) has argued that this apparent paradox results from not conceptualizing change as a separate dimension and that, once this is done in an item response model, evaluating the error in the ability estimates does not involve pretest and posttest correlations (= re-test reliability in classical test theory) so the apparent paradox disappears.

5(ii). Problems arising from the fact that, even on tests having little ceiling effect and constructed using IRT, individual change scores are highly negatively correlated with initial score.

Lord (1963) and Embretson (1991) have shown that the correlation between initial ability and change scores is necessarily negative, and therefore misleading. This arises from statistical phenomenon of regression to the mean. People who score below their true score on the pretest – perhaps because they are ill – have a positive change score. People who, for a similar reason, score below their true score on the re-test have a negative change score. Thus the scores of those scoring below the mean tend to go up and those scoring above the mean tend to go down – so the scores of the low ability group automatically go up! That is, the correlation between initial score and change score tends to be negative. This effect can be reduced by maximizing measurement precision at each level; e.g., by using IRT Tests, adaptive tests or by two independent measurements with the same test at time point 1; that is, the first measurement is used to describe the initial ability and the second measurement is used to calculate the gain score to the second time point (Fischer, 1974).

5(iii). Problems arising from the fact that differences, while equal in terms of the construct validity of the underlying measures, may have very different meaning or significance at different points in a scale.

We may take the high-jump as an example. For a 20 year old athlete who is 2 meters tall, to increase performance from 180 to 185 cm would not be a great challenge. Probably two hours of practice would suffice to bring it about. But it would be a completely different matter near the

maximum a motivated athlete is able to reach. Here a change of 5 cm in the height of the bar that can be cleared (e.g. from 220 cm to 225 cm) is a very big increase in performance, although the increase – 5 cm – is the same on the perfect Rasch scale used to measure the height of the bar.

Part II: A Way Forward – The Methodology Developed By Fischer

Having demonstrated that the problem of assessing change and, especially, differential change at different levels of ability, is fraught with difficulties, we now move on to review what can be done. The methodology to be described was developed by Fischer in Vienna (Fischer, 1972, 1974, 1983, 1995) in response to articles by such authors as Bereiter (1963), Cronbach & Furby (1970), and Holtzmann (1963). Although some of the authors just mentioned worked on group differences and others (e.g., Klauer, 1991, Liou, 1993, Ponocny, 2000) worked on individual change, we concentrate on Fischer's work because it seems to us that it is the most flexible, generalizable, and available for general use.

There are two main areas of application of the methodology which has been developed:

1. In the measurement and statistical assessment of change in *groups* (a) over time, (b) in response to different types or dosages of treatment(s), (c) in response to the same treatment(s) at different levels of ability, and (d) between groups differing in personality traits, gender, age, or any other observable characteristics.
2. To assess and compare change in *individuals.* Here one may want to know (a) how is a single individual changing over time, and, perhaps, to compare him or her with someone else; (b) to compare one person's response to different types or amounts (e.g. dosages) of treatment and then, perhaps, to compare those changes with those of other people having similar or different initial ability; and (c) to compare the responses of two or more people with different abilities to the same treatment.

1. The measurement and statistical assessment of change in groups

It is easiest to illustrate the principle on which Fischer's methodology is based by discussing a situation in which it is desired to document the differential effect of an experimental treatment on high and low ability

respondents (as in our example of average vs. superstar managers) although, as we shall shortly see, application of the method is by no means limited to such situations.

When the same test has been employed to assess performance before and after an intervention, each *item* that has been presented on the two occasions can be treated as if it were a pair of items with *different* item parameters within a Rasch scale**,** that is**,** as if it were a 'miniature Rasch scale' of length 2. For example, if one presents the same 10 items at pretest and posttest, one thereby obtains 10 miniature Rasch scales. There is no requirement that these items measure a common dimension; they could indeed be, and, in clinical studies, often are, actively chosen to measure 10 *different* dimensions in order to monitor change as *comprehensively* as possible. There is no need to use long tests, because each item measures a different latent dimension. (These dimensions may be correlated, or in some other way mutually dependent, or independent.)

After this one can, *in a second step*, assess whether any effects detected generalize across all items. If they do (and, from the many studies available, it would seem that it is indeed often the case), one can estimate an overall effect size for the treatment(s), or otherwise assess the relative effect sizes on the different "dimensions" involved. Obviously, the result is a very flexible set of procedures.

The same procedures can be applied to identify which *people* have changed. The general model assumes for each person the same effect on every 'miniature Rasch scale'. The model can be extended to identify clusters of people who are similar to others in the same group (in the sense that they respond more strongly to the experimental variables) but who differ from those in other groups who react in different ways. The procedures exactly parallel those used to identify clusters of *items* which behave like others within their groups but differ from those in other clusters.

Although the development of these procedures is formally grounded in IRT, the method mentioned departs fundamentally from the unidimensionality assumption of most IRT models. Because of this, the present model of change has been termed the "Linear Logistic Model with Relaxed Assumptions" (see Fischer, 1995b). The software required to implement it has been published By Fischer & Ponocny-Seliger (1998). Variations and extensions of the method to items with more than two ordered response categories are also available. (Readers interested in the psychometric background to this approach should refer to Fischer &

Molenaar (Eds.) (1995) and Fischer & Ponocny-Seliger (1998). However a short formal description of the method will be found in Endnote 4.)

These methods can be incorporated into various types of study design:

(1) Presentation of the same item sets at two or more time points to the same people. The items may, but need not, belong to an IRT model.

(2) Presentation of different, possibly overlapping, item samples from a unidimensional item pool (as established in a previous IRT study), at two or more time points. (More specifically, this design permits one to use, in one of the cases noted above, a more difficult test at posttest compared with pretest.) One or more unidimensional item pools may be used within the same study, so that the total item pool again becomes multidimensional. In such cases it is important that, at each time point, at least one item is selected from each unidimensional item pool, assuring that the relevant latent dimensions are actually measured at each time point. In principle, there is no limitation (other than test length) to the number of latent dimensions that can be included.

(3) The items may be dichotomous (as in most ability tests) or polytomous (with ordered response categories, as in many clinical rating scales).

(4) There may be any number of treatment and control groups. A treatment group is, by definition, a group of persons responding to the same subsets of items at the same time points and receiving the same treatment or treatment combinations.

(5) The data may be complete or incomplete.

Obviously, the range of admissible research designs is large. Given that a study is designed meaningfully with respect to the realized treatment combinations, the application of the methodology will yield estimates of effect parameters for both the treatments and other, possibly contaminating effects (such as simple aging), operating at the same time. The method also yields significance tests and standard errors for the effect parameters. Moreover, the software supports the formulation and testing of a number of standard hypotheses (e.g., generalizability of treatment effects or of amounts of change over both items subsets and person subgroups) as well as of a host of customized hypotheses.

At this point, because use will be made of it later, it is desirable to explain the conceptual shift that makes it possible to use IRT to solve these

hitherto intractable problems. One fundamental conceptual rearrangement has been to use a shift in *item* parameters (which generated the miniature Rasch scales in our previous discussion) as an index of change within *persons.* Technically speaking, the same item presented to respondents at two time points is formally considered as a pair of "virtual" items with different item parameters. The difference between the item parameters within pairs becomes an indicator of change in the respondents on the latent dimension behind them. Under the assumption of generalizability of change over the latent dimensions measured by different items and over persons within a treatment group, each pair of virtual items contributes to the overall information on the amount of change in that group. Therefore, combining all these contributions enables a measurement and statistical evaluation of change.

The estimation of effect parameters does not involve the estimation of item or person parameters. Only *change* parameters (i.e. the effects of treatment or changes which have occurred over time) are estimated. The computation is based entirely on those response combinations where a person has solved *only one* of the items of an item pair (= miniature Rasch scale). Response combinations where both responses to the items of a pair have been correct or both incorrect, provide no information on change and are therefore ignored. That is, it is advantageous to maximize the numbers of scores 1 (and neither 0 or 2) on each of these miniature Rasch scales (item pairs). This can be achieved by an intelligent selection of the items forming the pairs mentioned.

2. The measurement and statistical assessment of changes in individuals: (a) Over time; (b) Differentially in response to similar treatments; (c) Differentially in response to different treatments, and (d) For people having different patterns of ability and personality

The motivation to study change in individuals is usually different from that leading researchers to study change in groups: clinical psychologists ask whether an individual patient has been able, after a treatment period, to significantly improve his or her test performance level; educational psychologists want to compare individual growth within a certain time period to the average growth of the cohort; and applied psychologists may be interested in assessing the extent to which *an individual* has changed as a result of involvement in a training or personality development program. Perhaps more importantly, a coach, doctor, or teacher may want to

identify the specific training program, or combination of drugs, that is best for (i.e. produces the greatest change in) *a particular individual*

A first attempt at the measurement of change using IRT at the individual level was Embretson's (1991) "Multidimensional Rasch Model for Learning and Change". However, in this 'new model' only the simple difference score of the two person parameters (similar to the 'simple' difference between two raw scores), estimated by means of IRT at time point 1 and 2, was calculated. Fischer argued that the method was based on the asymptotic distribution of the person parameter estimations and that this requires lengthy testing procedures. Instead he suggested that only the change parameters themselves need be estimated.

The tests used could be achievement tests with dichotomous items or involve "Likert" type items having several (ordered) categories, like "always", "mostly", "rarely", and "never".

It has to be stressed, however, that – in contrast to group-oriented studies – the item pool used in this kind of study must be unidimensional because, if a study focuses on individuals, and if each item possibly measures a different dimension, only two discrete responses are available per latent dimension. This renders a scientific assessment of the amount of change on each latent continuum impossible. Besides this restriction to a unidimensional item pool, the present methodology has so far been developed only for two time points. In studies with more than two time points, pairs of item must be analyzed separately.

On the other hand, there is a great flexibility with respect to the composition of the tests used at the two time points: from any given unidimensional item pool it is possible to select any subset of items for use at each time point. Therefore, the respondent may be given the same items twice, or entirely different subsets of items may be selected for the pre and posttest, or the two items sets may overlap partially. If the researcher expects, for instance, an increase in the respondent's score on the ability or trait measured, he or she may choose easier items for the pretest than for the posttest, so that the expected shift on the latent dimension is roughly compensated for by an increase of item difficulty.

The idea behind the psychometric method is that the amount of change in the individual is projected onto change in the item parameters. The concept of "virtual" items thus again turns out to be essential for understanding the approach. Instead of thinking in terms of a change of the person (ability) parameter, it is helpful to imagine change as a shift of the posttest item parameters relative to the pretest item parameters.

Therefore, the person (ability) parameter – in spite of its change in reality – is technically considered as a constant, while the item parameters of the posttest items are exchanged for virtual item parameters. As a consequence, the responses given by the individual on both tests can be treated like responses of a respondent to just one test, the length of which is the sum of the lengths of the pretest and of the posttest. This makes it possible to employ the so-called "conditional maximum likelihood method". Its advantage is that the person parameter is eliminated from the further steps in the estimation and statistical testing procedures. Note that this approach avoids any asymptotic approximations since only the exact conditional distribution of the gain score is used.

This is not the place to describe the methodology in detail. Those who are interested should consult Fischer (1995a, 2000), although a brief formal discussion is given in Endnote 5. It should be mentioned, however, that partly similar methods have also been suggested by Klauer (1991), Liou (1993), and Ponocny (2000), in connection with the study of "person fit" in the Rasch model. Here it is sufficient to say that the method yields, for each individual, an estimate of the amount of change on the latent dimension, that this measure is independent of the true initial level of the trait or ability, that confidence intervals can be computed for the true individual mount of change, and that the amount of change can be tested for significance (see Endnote 6).

Examples of the use of the methodology we have discussed to overcome the intractable problems mentioned earlier will be found in Apfelthaler (2000), Erasim (1995), Fischer & Seliger (1997), Jenull-Schiefer (2000), Spiel & Glück (1998), Stögerer (2000), and Wernsdorf (1998). Here it is perhaps sufficient to illustrate the value of the approach by saying a little more about Prieler's (2000) previously mentioned study of the predictive validity of scores measuring response to stress in the Austrian army. A comprehensive battery of tests was administered to the officer cadets before and after the a strenuous night march with a view to discovering which *change* scores best predicted subsequent success as an officer.

The results of this study were as follows:

a. The predictive validity of both the pre- and posttest scores was low compared with the predictive validity of the change (gain or loss) scores, regardless of whether those change scores came from high, middle, or low scorers.
b. Three change (or difference) scores out of seven helped to predict those who would fail to complete the year-long officer training

program and thus yielded new, highly valid, criteria for personnel selection.

c. The predictive validity of some of these change scores was so great that one was entitled, on the basis of those scores alone, to drop all candidates with high negative change scores from the course.
d. Certain items had no predictive validity. As a result it was possible to shorten the test battery.

The Austrian Army has now adopted the method as a standard procedure. Both, the testing under load in connection with the new measurement approach guarantee a high personnel quality management, which reduces personnel costs to an enormous amount and enlarge the motivation of each soldier in a considerable way.

In his research in the German army, Melter (1992) came to the conclusion that 'already in the in selection phase, there are minimal information about a later break-off of a soldier career'. Using the procedures that have been described during the initial phase of recruitment it is now easier to detect soldiers who might break off their careers later. In other words, it seems that more value is now being obtained from the existing data than was previously the case.

Summary

We have described a series of problems, of increasing complexity, that have bedeviled psychological research and, in some cases, led researchers to draw seriously misleading conclusions that have in turn had major detrimental, indeed unethical, effects on policy.

Some of these problems have been widely recognized since the birth of statistics even if, as the APA task force on statistical inference belatedly noted, being widely overlooked by research teams anxious to apply "sophisticated" statistical procedures to their data without first becoming thoroughly familiar with it.

Some of the other problems we have discussed have been recognized by small groups of methodologists for perhaps 40 years. However, even when they were recognized, no readily applicable solutions were available. As a result, most theoretically-oriented researchers, while sometimes troubled by a vague awareness of problematic features of their methods, felt that they had to do *something* and continued to utilize simplistic

procedures unaware of just how seriously the measurement errors that were involved undermined any prospects of arriving at meaningful conclusions. To these theoretically-oriented researchers has been added a vast army of researchers charged with the practical task of demonstrating the differential value of different dosages or different drugs, comparing the cost-effectiveness of different therapies for different "kinds of people" within managed health care programs, or determining which of a number of therapeutic regimes (with different costs) was conferring the greatest benefit on an individual patient.

Yet even this range of methodological problems does not cover those encompassed within this article. Other serious problems stem from the fact that most of the evaluations currently conducted – whether at an individual or group level – are insufficiently comprehensive and thus lead to misleading conclusions and inappropriate action.

It is of interest that, while perhaps aware of these problems, the APA task force did not think the problems we have mentioned sufficiently serious to highlight to them.

Having described some of the problems that inhere in the measurement of change in groups and individuals, we summarized a new, ingenious, theoretically-based, practical, set of solutions to those problems. Even if new in no other way, these are new in the sense that the computer programs needed to implement them have only recently become available.

Two, perhaps serendipitous, benefits of these procedures may be singled out for final mention: They make it possible to find out whether a particular *individual* has responded in a significant way to some drug, therapy, or educational or developmental program (and, if so, in what way). And they make it more feasible to focus on comprehensiveness in the evaluation of people and programs by reducing the length of the tests that are deemed to be necessary.

All this having been said, it remains to add a word of warning. The procedures that have been described are no panacea. The simultaneous estimation of item and treatment effect parameters can be a considerable source of bias. One should, therefore, choose one of two approaches whenever possible: Either one should determine the item parameters in an independent calibration study (e.g. using IRT tests for which item difficulty parameters are available); or present the same items repeatedly.

Notes

7.1. More precisely, the raw score differences that correspond to equal differences in latent ability vary markedly with: (a) the general difficulty level of the test relative to the ability tested and (b) the distribution of the item parameters relative to the interval on the latent trait where change occurs (Fischer & Prieler, 2000), that is to say, with the section and shape of what Fischer (1991) has termed the 'test characteristic curve' at which the difference is measured. This is true even of tests which satisfy the criteria of the most popular versions of Item Response Theory (IRT) models – that is to say, it is true for tests which conform the 1-Parameter (1PL) model (i.e. a model which does not allow for variation in the slope of the ICCs or guessing), the 2-Parameter (2PL) model (i.e. a model which allows variation in slope but not guessing), and the 3-Parameter (3PL) (i.e. the one which allows for both guessing and variation in slope) variants of IRT (see Hambleton, Swaminathan, & Rogers, 1991, and van der Linden & Hambleton, 1997, for a fuller discussion of these models). Still more pointedly, note that unidimensionality of the kind which is assured by conformity to most IRT models (Hambleton, Swaminathan & Rogers, 1991 & van der Linden & Hambleton, 1997), does not in itself lead to bias-free difference scores.

7.2. Earlier discussions of some of these problems will be found in Bereiter, 1963; Harris, 1963; Lord & Novick, 1968; Cronbach & Furby, 1970; Williams & Zimmermann, 1996; Guthke, 1996; Rost, 1996

7.3. We may begin by quoting Fischer & Molenaar (1995):

> "In a psychological test or attitude scale, one tries to measure the extent to which a person possesses a certain property. The social and behavioral sciences often deal with properties like intelligence, arithmetic ability, neuroticism, political conservatism, or manual dexterity. It is easy to find examples of observable human behavior indicating that a person has more or less of such general property, but the concept has a surplus value, in the sense that no specific manifest behavior fully covers it. This is the reason why such properties are called latent traits. The use of a test or scale presupposes that one can indirectly infer a person's position on a latent trait from his or her responses to a set of well-chosen items, and that this also allows us to predict his or her behavior when confronted with the items from the same domain. A statistical model of the measurement process should allow us to make such predictions. Moreover, it should allow generalization to other persons taking the same test, and address the question of generalizability over test taking occasions.
>
> "The observed behavior are test scores, whereas, in classical theory, the independent variables are the latent variable and error. The dependent variable is the observed test score.
>
> Observed Score = True Score + Error

"However, this "True Score Theory" has several shortcomings for test construction (Hambleton, Swaminathan, & Rogers, 1991). First, item difficulty and item discrimination indices are group-dependent. A problem arises when the examinee sample does not closely reflect the population for whom the test is intended, and thus the usefulness of the statistical indices obtained in the sample will be limited. Second, examinee ability estimates, which are item dependent, rely on the particular choice of items selected for the test. This makes the comparison difficult when examinees take different tests. Third, the standard error of measurement is assumed to be the same for all examinees, which is implausible because scores on any single test are not equally precise measures for examinees of different ability. Finally, the reliability is defined as the correlation between test scores on "parallel" tests, which, in practice, are difficult to construct."

For this reason, psychometricians have sought alternative theories and models of measurement. Item Response Theory (IRT) overcame the limitations of classical test theory. In IRT the observed behavior is individual item responses (e.g., 0 or 1). The latent variable influences the probabilities of the responses to the items. The probability that a person will pass or endorse a particular item depends on their trait level and on the difficulty of the item, as follows:

Prob (Item passed) = Function [(Trait level) – (Item Difficulty)]

The Rasch IRT model is a simple logistic function of trait level and item difficulty, as follows:

$$P(X_{ij}=1/\theta_j, b_i) = \exp(\theta_{vi}-b_i)/1 + \exp(\theta_{vi}-b_i)$$

where θ_j is the person's trait level and b_i the item's difficulty level (Rasch, 1960; Embretson, 1999).

Although most authors would wish in principle to limit the use of the phrase "conforms to the Rasch model" to tests which conform to the above single-parameter logistic model, the heated debates that have raged about whether the Raven Progressive Matrices conform to "The Rasch Model" indicate that this is not always the case in practice. The sets of ICCs for the items of the RPM that have been drawn and published from a range of studies conducted in different countries since 1935 (See Raven, 1981) clearly show that the curves vary in slope and that a "guessing" parameter operates before the items start to discriminate. It follows that only a 3-parameter model will suffice and thus that a purist interpretation of the requirements for conforming to "the Rasch model" cannot be satisfied. Yet not only did Rasch himself test his model by showing that it fitted the RPM, other researchers (such as Andrich) who are steeped in both IRT and Rasch modeling have made this claim. It follows that the purist interpretation of what constitutes "the Rasch Model" has not always been adopted in practice. As Hambleton has been at pains to point out (without always being listened to), these heated debates revolve around two questions: (a) The nature and level of the variation in the *criteria* that have

been set for acceptance as "conforming to the Rasch model", and (b) the nature and size of the samples studied ... for it turns out that the item and test parameters one obtains from samples of the size typically studied by psychologists (which are also characterized by a very restricted range of scores) are very unstable, particularly in the case of 3-parameter models. In this paper we use the term "Rasch model" to refer to any test which conforms to the 1, 2, or 3-parameter model.

7.4. The following short formal description of the LLRA comes from Fischer & Ponocny-Seliger, (1998).

Consider a test comprising k Items I_i that are given to a set of persons S_v at two time points, T_1 and T_2, before and after certain treatments. Let the probabilities of positive responses '+' to item I_i be

$$P(+/S_v, I_i, T_1) = \exp(\theta_{vi})/1 + \exp(\theta_{vi}) \text{ at } T_1,$$

and

$$P(+/S_v, I_i, T_1) = \exp(\theta'_{vi})/1 + \exp(\theta'_{vi}) \text{ at } T_2.$$

Parameter θ_{vi} denotes S_v's position at time point T_1 on the latent dimension D_i which is measured by Item I_i. The θ_{vi} are allowed to vary freely, so the items may measure independent traits, or correlated or otherwise mutually dependent traits. Since, therefore, no restrictions are imposed on the admissible relations between the traits, the model is applicable to a wide variety of substantive domains. Similarly, θ'_{vi} is S_v's position on dimension D_i at time point T_2.

To get real valid results, the size of each treatment group should not be below 30 if the test is of a normal length (i.e. have more than 10 items). In fact, the validity of the result depends on the factor n* k. It follows that it is theoretically possible to have less than 30 people, but in this case it is necessary for the test to have an unreasonable number of items.

The disadvantages of the method is that an inconvenient result is reached if there is no generalizability over parts of items or persons at all (this means that the treatment effect is different for every item or person; a result which is normally not interpretable).

7.5. The following short, formal description of the LLTM for the assessment of change of individuals comes from Fischer (2001).

Let the item sample consist of items $I = I_1,\ldots,I_k$. It is assumed that the PCM (Partial Credit Model) fits these items:

$$P(X_{vij} = 1/\theta_{vi}, \beta_{i0},\ldots,\overset{mi}{\beta}_{imi}) = \exp(j\theta_v + \beta_{ij}) / \Sigma_{l=0} \exp(l\theta_v + \beta_{il})$$

Where:

X_{vij} denotes the random response variable with realizations $x_{vij} = 1$ if person S_v's response to item I_i belongs to response category C_{ij}, and $x_{vij} = 0$ otherwise;

m_i is the number of items I_i's response categories minus 1 (the response categories being numbered 0,...,m_i);

θ_v is testee S_v's person parameter;

β_{ij}, for j = 0,....,m_i, are item I_i's item x category parameters.

To make the model identifiable, $k + 1$ normalization conditions have to be imposed on the parameters. Note that RSM (Rating Scale Model) and RM (Rasch Model) are special cases of the PCM.

The model above is assumed to hold for all items if they are presented on a single occasion. Therefore, the model holds for the items of pretest I_1. For posttest I_2, however, the model has to be rewritten since it is assumed that the person parameter θ_v has to be replaced by $\theta_v + \eta_v$, where η_v is a parameter representing the amount of change in person S_v. The conditional maximum likelihood method (CML) is used to estimate the change parameters η_v.

So if there exists a test which fits a PCM (or RSM or RM=1PL, 2PL, 3PL), it is possible to calculate a table for significant changes between two time points on the level of an individual.

The disadvantage of the method is that for persons with perfect or zero scores on two time points you can say nothing about the real change on the latent trait.

7.6. One interesting extension of the LLRA/LLTM models for the assessment of change discussed here is that of Meiser (1995). Alternatives to the LPCMWin software, which is based on the Conditional Maximum Likelihood (CML) approach, are to be found in a software package based on the Marginal Maximum Likelihood (MML) approach from Wu, Adams, & Wilson (1995) and a software program based on a linear model for Log-Odds-Quotients (Linacre, 1996).

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

APA Task force on Statistical Inference (1999). See: L. Wilkinson and Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604.

Apfelthaler, E. (2000). Medikamentöse und psychotherapeutische Effekte in der Behandlung der erektilen Dysfunktion. Unpublished dissertation, University of Vienna.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 3–20). Milwaukee, WI: University of Wisconsin Press.

Bouvier, U. (1969). *Evolution des Cotes a Quelques Tests.* Belgium: Centre de Recherches, Forces Armees Belges.

Budoff, M. (1973). Measuring learning potentials: An alternative to the traditional intelligence test. *Studies in Learning Potentials, 3,* 39ff.

Budoff, M., Corman, L., & Gimon, A. (1976). An educational test of learning potential assessment with Spanish speaking youth. *Inter-American Journal of Psychology, 10,* 13–24.

Carver, R. P. (1989). Measuring intellectual growth and decline. *Psychological Assessment, 1(3),* 175–180.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change" – or should we? *Psychological Bulletin, 74*, 68–80.

Embretson, S. E (1991). Implications of a multidimensional latent trait model for measuring change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change.* Washington, DC: American Psychological Association.

Erasim, U. (1995). Anwendung des LLRA: "Der Einfluss von mentalem Training auf die sportliche Leistung jugendlicher Tennisspieler" . Unpublished dissertation, University of Vienna.

Feuerstein, R. (1979). *The dynamic assessment of retarded performers*. Baltimore: University Park Press.

Feuerstein, R., Klein, P., & Tannenbaum, A. (Eds.). (1990). *Mediated learning experience: Theoretical, psycho-social, and educational implications*. Proceedings of the First International Conference on Mediated Learning Experience. Tel Aviv: Freund.

Fischer, G. H. (1972). A measurement model for the effect of mass-media. *Acta Psychologica, 36,* 207–220.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* (Introduction to mental test theory). Bern: Huber.

Fischer, G. H. (1977). Linear logistic latent trait models: Theory and application. In H. Spada & W. F. Kempf (Hrsg.). *Structural models of thinking and learning.* Bern: Huber.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 46*, 59–77

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika, 54*, 599–624.

Fischer, G. H. (1991). A new methodology for the assessment of treatment effects. *Evaluación Psicológica/Psychological Assessment, 7(2),* 117–147.

Fischer, G. H. (1995). Some neglected problems in IRT. *Psychometrika, 60 (4),* 459–487.

Fischer, G. H. (1995). Linear logistic models for change. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models, recent developments and applications* (pp. 158–180). New York: Springer-Verlag.

Fischer, G. H. (1997). Multidimensional linear logistic models for change. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 323–346). New York: Springer-Verlag.

Fischer, G. H. (2001). Gain scores revisited under an IRT perspective. In A. Boomsma, M. A. J. van Dujin, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 43–68). New York: Springer-Verlag.

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models, recent developments and applications.* New York: Springer-Verlag.

Fischer, G. H., & Seliger, E. (1997). Multidimensional linear logistic models for change. In: W. J. van der Linden & R. K. Hambleton, *Handbook of modern item response theory* (pp. 323–346). New York: Springer.

Fischer, G. H., & Ponocny-Seliger, E. (1998). *Structural Rasch modeling.* Handbook of the Usage of LPCM-Win 1.0, ProGAMMA

Fischer, G. H., & Prieler, J. A. (2000). *An IRT-based methodology for the assessment of change.* Appendix 2 in J. Raven, J. C. Raven, & J. H. Court, *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3*: *The Standard Progressive Matrices*. Oxford, England: Oxford Psychologists Press; San Antonio, TX: The Psychological Corporation.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95,* 29–51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171–191.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54(1),* 5–20.

Flynn, J. R. (2000). IQ gains, WISC subtests and fluid ***g***: ***g*** theory and the relevance of Spearman's hypothesis to race. In G. R. Bock, J. A. Goode, & K. Webb (Eds.), *The Nature of intelligence* (pp. 202–227): *Novartis Foundation Symposium 233.* Chichester, UK: Wiley.

Guthke, J. (1982). The learning test concept – An alternative to the traditional static intelligence test. *German Journal of Psychology, 6,* 306–324.

Guthke, J., & K. H. Wiedl (1996). *Dynamisches Testen.* Gottingen: Hogrefe Verlag.

Hambleton, R. K. (1988). Comments made in a symposium on *Questionable Assumptions in Test Construction*, held at the meeting of the International Association for Applied Psychology, Sydney.

Hambleton, R. K. (1989). Constructing tests with item response models: A discussion of methods and two problems. *Bulletin of the International Test Commission, No.28/29*, 96–106. Strasbourg: ITC.

Hambleton, R. K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press.

Harris, C. W. (1963). *Problems in measuring change.* Madison, Milwaukee, WI: University of Wisconsin Press.

Jenull-Schiefer, E. (2000). Evaluation verhaltenstherapeutischer Gruppentherapieprogramme zur Verbesserung sozialer Fertigkeiten bei schizophren Erkrankten Unpublished doctoral dissertation, University of Vienna.

Holtzmann, W. H. (1963). Statistical models for the study of change in the single case. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 199–211). Milwaukee, WI: University of Wisconsin Press.

Klauer, K. C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika, 56,* 213–228.

Linacre, J. M. (1996). Generalizability theory and many-facet Rasch measurement. In G. Engelhard, jr. & M. Wilson (Eds.), Objective measurement. Theory into practice, (Vol. 3, pp. 85–98). Norwood, NJ: Ablex Publishing Corporation.

Liou, M. (1993). Exact person tests for assessing model-data fit in the Rasch model. *Applied Psychological Measurement, 17,* 187–195.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change.* Madison: University of Wisconsin Press.

Lord, F. M., & Novick, M.R. (Eds.). (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Meiser, T. (1996). Loglinear Rasch models for the analysis of stability and change. *Psychometrika, 61,* 629–645.

Melter, A. H. (1992). Psychologische Untersuchung der Offiziersanwärter mit nicht erfolgreicher Offiziersausbildung. In M. Rauch (Hrsg.), Jahrbuch des psychologischen Dienstes der Bundeswehr, München: Verlag für Wehrwissenschaften.

Molenaar, I. W. (1995). Some Background for Item Response Theory and the Rasch model. In: G. H. Fischer & I. W. Molenaar (Eds.), (1995). *Rasch models, recent developments and applications* (pp. 3–14). New York: Springer-Verlag.

Ponocny, I. (2000). Exact person fit indexes for the Rasch model for arbitrary alternatives. *Psychometrika, 65,* 75–106.

Prieler, J. A. (1998, July). *Validation of Personnel Selection in the Austrian Army*. Paper presented at the International Congress of Applied Psychology, San Francisco.

Prieler, J. A. (2000). *Evaluation eines Ausleseverfahrens für Unteroffiziere beim Österreichischen Bundesheer* (Validation of personnel selection of officers in the Austrian Army). Unpublished doctoral dissertation, University of Vienna.

Rasch, G. (1947). Quoted by B. D. Wright in a foreword to Rasch, G. (1980). *Probabalistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Illinois Press.

Raven, J. (1981). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.1: The 1979 British Standardisation of the Standard Progressive Matrices and Mill Hill Vocabulary Scales, Together With Comparative Data From Earlier Studies in the UK, US, Canada, Germany and Ireland.* Oxford,

England: Oxford Psychologists Press; San Antonio, TX: The Psychological Corporation.

Raven, J. (1989). Questionable assumptions in test construction. *Bulletin of the International Test Commission, 28 & 29,* 67–95.

Raven, J. (1991). *The tragic illusion*: *Educational testing*. New York: Trillium Press.

Raven, J. (1994). *Managing education for effective schooling*: *The most important problem is to come to terms with values*. Unionville, New York: Trillium Press.

Raven, J. (1995). *The new wealth of nations*: *A new enquiry into the nature and origins of the wealth of nations and the societal learning arrangements needed for a sustainable society*. Unionville, New York: Royal Fireworks Press; Sudbury, Suffolk: Bloomfield Books.

Raven, J. (2000a). Ethical dilemmas. *The Psychologist, 13,* 404–406.

Raven, J. (2000b). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41,* 1–48.

Raven, J., Raven, J. C., & Court, J. H. (1998a). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1*: *General Overview.* Oxford, England: Oxford Psychologists Press; San Antonio, TX: The Psychological Corporation.

Raven, J., Raven, J. C., & Court, J. H. (1998b). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 4*: *The Advanced Progressive Matrices.* Oxford, England: Oxford Psychologists Press; San Antonio, TX: The Psychological Corporation.

Raven, J., Raven, J. C., & Court, J. H. (2000). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3*: *The Standard Progressive Matrices.* Oxford, England: Oxford Psychologists Press; San Antonio, TX: The Psychological Corporation.

Raven, J., & Stephenson, J. (Eds.). (2001). *Competence in the learning society.* New York: Peter Lang.

Raven, J. C. (1965). *Advanced Progressive Matrices, Sets I and II: plan and use of the scale with a report of experimental work carried out by G. A. Foulds & A. R. Forbes.* London: H. K. Lewis.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion.* Bern: Huber.

Schaie, K. W., & Willis, S. L. (1986). *Adult development and ageing* (2nd edition). Boston: Little Brown.

Spearman, C. (1926). *Some issues in the theory of **g** (Including the law of diminishing returns)*. Address to the British Association Section J – Psychology, Southampton, England, 1925. Bound in Collected Papers, Psychological Laboratory, University College of London.

Spiel, C., & Glück, J. (1998). Item response models for assessing change in dichotomous items. *International Journal of Behavioral Development, 22(3),* 517–536.

Stögerer, P. (2000). Prognostische Bedeutung von "Angst vor Kontrollverlust" für die Therapie des Paniksyndroms Unpublished dissertation, University of Vienna.

Teasdale, T. W., & Owen, D. R. (1989). Continuing secular increases in intelligence and a stable prevalence of high intelligence levels. *Intelligence, 13,* 255–262.

Thorndike, R. L. (1975). *Mr. Binet's Test 70 years later.* Presidential Address to the American Educational Research Association.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.

Wernsdorf, T. (1998). Konzentrative Bewegungstherapie und Ich-Erleben. Unpublished dissertation, University of Vienna.

Williams, R. H., & Zimmermann, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement, 20,* 55–69.

Wright. B. D. (1968). Sample free test calibration and person measurement. *Proceedings of the 1967 invitational conference on testing problems.* Princeton, NJ: Educational Testing Service.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*: *Rasch measurement.* Chicago: Mesa Press.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1995). *MATS*: *Multi-aspect test software computer program.* Melbourne: Australian Council for Educational Research.

PART III

Stability and Change in RPM Norms Across Time And Cultures

The first chapter in this Part of our book begins by summarising research into the stability and change in RPM norms across culture and time that had mainly been conducted prior to the year 2000. This leads to a fairly detailed discussion of the environmental factors that do, and do not, influence RPM scores.

In the next chapter, Francis van Dam reports the results of a *longitudinal* study of change and stability in Advanced Progressive Matrices scores from approximately 20 to 50 years of age. Contrary to many people's expectations, there is only a slight decline in average scores with increasing age. This confirms Flynn's hypothesis that the cross sectional data which had previously been thought to show a decline with age mainly reflects an intergenerational increase in scores with date of birth. But the most striking result of the study is the extent to which some people's scores had actually *increased.*

Other chapters report the results of norming studies recently conducted in Slovenia, Lithuania, Turkey, Kuwait, South Africa, and Tribal areas of India. These data, together with those in the chapter summarising the results of the Romanian study, extend the data base of cross-cultural norms summarised in the *General Introductory* chapter to this book and those in the initial chapter of this Part of it. Once again, while the similarity and differences between the norms for these different groups, and the variance within each of them, is striking, the explanation of the variance within and between groups proves as elusive as ever, particularly as it is yet again shown not to be due to variance in the scaleability of the measure.

Chapter 8

Change and Stability in RPM Scores Over Culture and Time: The Story at the Turn of the Century

John Raven

Introduction

This chapter is based upon a paper published in *Cognitive Psychology* (Raven, 2000b), which was itself based on material previously published in the *Manuals* for the *Raven Progressive Matrices* – see various reference entries for Raven, Raven, and Court. It offers a more extended summary of research relating to the stability and change in RPM scores over culture and time (and their causes) than was included in the *General Introduction* to this book.

Given that the tests have been in use for 70 years, distilling off the evidence bearing on the question of how similar are the norms for different cultural groups and how have they changed over time is not so easy as might be expected. The reasons why the data are not better than they are merit review because of their implications for future work in the area.

One reason why the task is so difficult is that, as Dahlstrom (1993) noted in an article appropriately titled "*Tests: Small samples, large consequences"* that most of the studies in the area are not only of poor quality but also conducted for other purposes. Thus most of the research in which the RPM have been used have sought to relate RPM scores to some other variable (such as educational or occupational performance) rather than to assemble basic normative data. And, both in these studies and in those which attempted to provide some kind of reference data, the researchers concerned have been relatively indifferent to the importance of sampling … even though this has major implications for the validity of the significance tests they sought to apply. Thus many researchers have

tried to *explain* apparent cultural differences without first establishing just how large or pervasive those differences were.

Creeping Awareness of the Importance of Studying Change Over Time

However, so far as change over time is concerned, one reason why there is little adequate data is that, since most psychologists never even suspected that scores would increase over time, they not only did not think it was necessary to collect data which would bear on the question, they simply assumed that normative data collected in the past were still applicable. In short, they did not see any need to restandardise tests.

Another factor is, however, that the available evidence that scores were increasing over time was misinterpreted – as were the similar data available from numerous cross-sectional studies conducted with other tests – as evidence of declining ability after 20 years of age.

Although some researchers (e.g. Owens, 1966; Bouvier, 1969; Thorndike, 1975) did notice an apparent increase in scores on some components of "intelligence", the overall effect, when scores on these sub-tests were combined with the others, could hardly be described as "dramatic". Had the researchers concerned been in the habit of thinking in Spearman's terms – separating eductive and reproductive abilities in their minds, they might have noticed that it was essentially only one component of "intelligence" (namely, *eductive* ability) which was increasing and that the increase in such scores was indeed dramatic. (In this context we may note that, besides, in 1984 and 1987, documenting the rise more thoroughly than previous workers, Flynn's main contributions (eg Flynn, 2000) have in fact been to draw attention, first, to the *rate* of change in eductive ability, and, second, by forcefully raising the question of whether this increase has resulted in a genuine increase in *knowledge*, to underline the *differential* rate of change in eductive and reproductive abilities.)

Beyond these background constraints, there were also substantial methodological problems. First, to be meaningful, the data had to be sectioned by age, as in Table 8.1 (which will be explained more fully later). Second, the bimodal and skewed within-age distributions shown in Figure 8.1 (redrawn from J. Raven, 1981), combined with a scatter which varied with age (also illustrated in Figure 8.1), meant that the usual data reduction techniques (i.e. reduction of the data to means and standard deviations) could not meaningfully be adopted.

Nevertheless, despite the validity of most of these components in an explanation of psychologists' failure to notice the increase in scores that

Table 8.1. *Standard Progressive Matrices*
1979 British Percentile Norms for the Self-Administered or Group Test Among Young People (Smoothed)

	Age in years (months)																		
	6½	7	7½	8	8½	9	9½	10	10½	11	11½	12	12½	13	13½	14	14½	15	15½
	6(3)	6(9)	7(3)	7(9)	8(3)	8(9)	9(3)	9(9)	10(3)	10(9)	11(3)	11(9)	12(3)	12(9)	13(3)	13(9)	14(3)	14(9)	15(3)
	to	to	to	to	to	to	to	to	to	to	to	to	to	to	to	to	to	to	to
Percentile	6(8)	7(2)	7(8)	8(2)	8(8)	9(2)	9(8)	10(2)	10(8)	11(2)	11(8)	12(2)	12(8)	13(2)	13(8)	14(2)	14(8)	15(2)	15(8)
95	33	34	37	40	42	44	46	48	49	50	51	52	53	54	54	55	56	57	57
90	30	32	35	38	40	42	44	46	47	48	49	50	51	52	53	54	54	55	55
75	22	26	30	33	36	38	41	42	43	44	45	46	47	49	49	50	50	51	51
50	**16**	**19**	**22**	**25**	**31**	**33**	**36**	**38**	**39**	**40**	**41**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**47**
25	13	14	15	17	22	25	28	32	33	34	36	37	38	39	41	42	42	42	42
10	10	12	12	14	16	17	19	23	27	29	31	31	32	33	35	36	36	36	36
5	9	10	11	12	13	14	15	17	22	24	25	26	27	28	29	30	33	33	33
n	112	138	148	174	153	166	198	172	194	187	164	164	174	185	180	196	189	191	171

Note. Based on a nationally representative sample of British schoolchildren, excluding those attending special schools (see Raven, 1981 for details). Younger and less able children were tested individually.

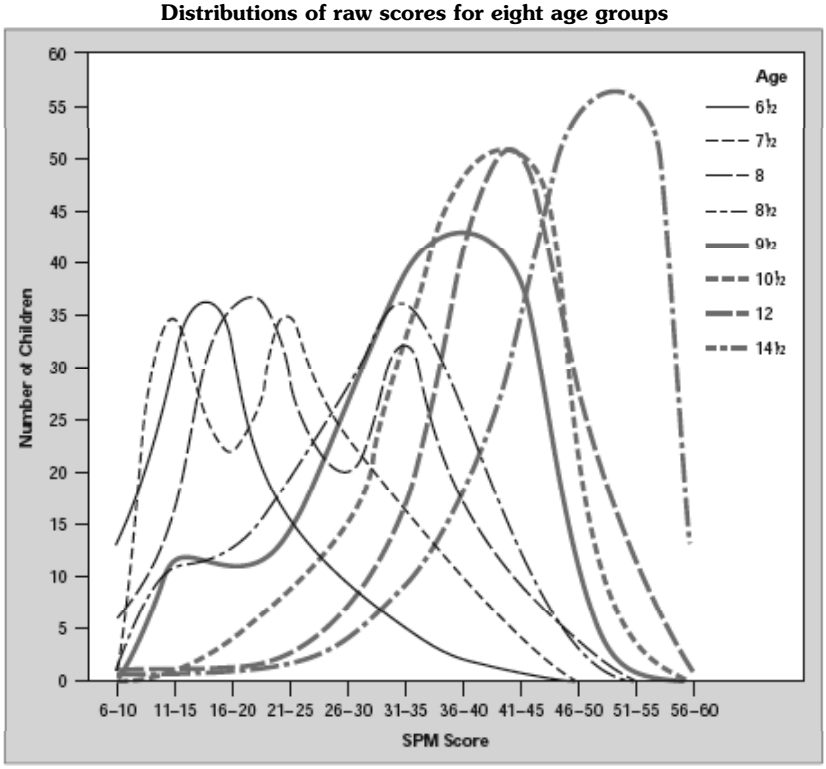
was occurring, the most important is probably that, until 1979, such data as were available would not really have led anyone to suspect an increase – even on *Progressive Matrices.* So let us now review such data as were available to shed light on changes over time and culture for young people and later do the same thing for the adult data.

Studies of Young People

1. Studies relating to change over time

The *Standard Progressive Matrices* (SPM) was first fully standardised by J.C. Raven on 1,407 children in Ipswich, England, in 1938 (J. C. Raven, 1941). The next substantial study (J. C. Raven & Walshaw, 1944) was conducted, not in order to produce norms for the RPM, but to gather equivalent data for the *Mill Hill Vocabulary Scale* (MHV). It was carried

Figure 8.1. *Standard Progressive Matrices*
1979 British Standardisation
Distributions of raw scores for eight age groups



out in a town not far from Ipswich, namely, Colchester, in 1943-44. The SPM norms obtained in that study were consistently two raw score points *lower* than the Ipswich norms. In 1952, Adams reported norms from 11,621 12-year-old children in Surrey, England. These data were, within the limits of sampling error, very similar to Raven's 1938 (Ipswich) norms. Tuddenham, Davis, Davison, and Schindler (1958), in one of the few studies which attempted to establish the appropriateness or otherwise of the British norms in the United States, tested several school classes of Californian children. They concluded that the British norms were acceptable. In 1963-65 Skanes tested 4,017 children aged 9 ½ to 14 years in St. John's, Newfoundland. The similarity between Skanes' results and the 1938 Ipswich norms is striking (J. Raven, 1981). Later, in 1967, in Corner Brook, Newfoundland Skanes tested the entire population (2,097) of children aged 10 ½ to 14 ½ years. The results consistently lagged *behind* the Ipswich norms. In 1972, Byrt and Gill (1973), working with the author, collected data from a nationally representative sample of 3,464 primary school children aged 5 ½ to 11 ½ years in the Republic of Ireland. The urban norms seemed to corresponded to the 1938 Ipswich norms, although the figures for rural areas lagged behind.

As late as 1979 – 40 years after the test was published – therefore, there was little to suggest a secular increase in scores. Quite the contrary: everything suggested stability.

From 1979 onwards the story began to change. In that year, Kratzmeier and Horn (1979) reported norms from a large German study which were well above those obtained in England in 1938. Melhorn's (1980) East German data were similar. The 1979 British norms, compiled (with the aid of a Social Science Research Council grant and assistance from the Government Office of Population Censuses and Surveys) from a carefully drawn sample designed to represent both the whole of Great Britain and the socioeconomic variance within it, appeared to be broadly similar to those obtained in the two German studies (J. Raven, 1981). Holmes (1980) reported results for British Columbia (Canada) which were similar to, if slightly lower than, the 1979 U.K. national norms. Both the Australian Council for Educational Research (see de Lemos, 1984, 1989) and the New Zealand Council for Educational Research (1984) reported closely corresponding results for their respective countries. Ferjencik (1985) reported data for the *Coloured Progressive Matrices* (CPM) for what was then Czechoslovakia which corresponded to a recently reported British study. Work carried out in the U.S. by J. Raven (2000a)

revealed that, while the overall U.S. norms lagged behind these new international figures, the White norms did not. Zhang and Wang (1989) collected data for urban mainland China which showed that, despite what had been suggested by the high norms reported by Chan (1981, 1989) for Hong Kong, norms for a sample designed to be representative of urban mainland China corresponded closely to recent norms obtained elsewhere. More recently, similar data have been reported for Poland (Jaworowska & Szustrowa, 1991; J. Raven, J.C. Raven & Court, 1998c, 2000, updated 2004), Spain (J.C. Raven, Court, & J. Raven, 1995), further school districts of the U.S. (J. Raven & Court, 1989), Switzerland (Martinolli, 1990; Spicher, 1993) and India (Deshpande, in J. Raven , J.C. Raven, & Court, 2000).

Two observations may be interjected at this point. First, when reporting the results of the 1979 British standardisation, we ourselves (J. Raven, 1981), while noting the difference between the 1938 and 1979 norms, failed to comment on its *magnitude* and, overlooking the fact that the scores of the more able adolescents approached the maximum obtainable on the test, suggested that the increase had mainly occurred among the less able. Second, given the similarity in the norms reported by all the researchers listed in the last paragraph who published data from 1936 to 1979 and the similarity in the 1980s norms reported by the other authors whose work has been summarised, there was no hint that we might be looking at evidence of a *continuous* increase in scores over time. There could just have been a jump.

2. Geographical and cultural variance

The studies outlined thus far suggest that the norms for different populations are similar at a given point in time but had somehow jumped dramatically in the 1970s.

We will now summarise studies documenting variance in the norms for young people from different geographical areas and between cultures both as a topic in its own right and with a view to exploring what light they are able to shed on the changes over time. Studies revealing broad differences between countries will be reviewed first and followed by a review of studies of variance within countries.

As has been mentioned, Chan's Hong Kong norms exceed most of the norms already mentioned. However the norms which most significantly exceed them come from Taiwan (Miao & Huang, 1990; Miao, 1993). (To reduce the likelihood of an inappropriate interpretation being placed

on these norms it should, be mentioned that, in the course of a visit to Taipai, it emerged that not only does the RPM play a major role in the rigid Taiwanese system of educational assessment and school placement, all teachers are supplied with copies of the test and encouraged to "train pupils in its use". It therefore seems likely that the "high" Taiwanese norms reveal, not the superior eductive ability of the Taiwanese, but rather, how small is the maximum effect of motivation and training.)

On the other hand, as also noted, norms for rural and isolated communities are typically lower than others. The previously mentioned norms for the Republic of Ireland and Newfoundland can, in this context, be seen to confirm this. Other low norms for what appear to be good samples of the relevant populations have been reported for Brazil (Angelini, Alves, Custodio, & Duarte, 1988), Turkey (Sahin & Duzen, 1994), Malaysia (Chiam, 1994, 1995), Puerto Rico (Kahn, Spears, & Rivera, 1977; J. Raven & Court, 1989), and a remote area in the mountains of Peru (see Munoz in Raven et al., 1998b).

As emphasised by J. Raven (1989), the "low" norms reported in most of these studies must be set in an appropriate context by observing that, with the notable exception of the Peruvian mountain norms, most are above the British 1938 norms. It follows that the factors that have been responsible for the shortly-to-be-discussed increases in scores over time could also have caused the differences between cultural groups.

More systematic studies of the variance between geographical, socioeconomic, and ethnic groups within countries were undertaken in the course of both the British and U.S. standardisations among young people. Because both the designs and the variables considered in these two studies were different they must be discussed separately.

The 1979 British standardisation

The 1979 British Standardisation was conducted in seven areas of the country which were chosen, under the guidance of the Government Office of Population Censuses and Surveys (OPCS), to represent all the types of area into which a cluster analysis of large amounts of demographic data had shown the socio-economic variance within the country could be classified (Webber, 1977). The types of area in which few people lived were over-sampled in order to have enough respondents to make it possible to break the data down by type of region. Later, the data were re-weighted to its correct proportions to give overall statistics. It was therefore possible to employ fairly sophisticated statistical procedures when analysing the data. Altogether 3,250 children aged 6 to 16 were tested.

This is a convenient point at which to explain the format in which the data will be displayed. Many authors present their data in terms of Deviation IQs with a mean of 100 and a standard deviation of 15. This process is, in general, unjustifiable for a number of reasons which include two that are important here: First, as was evident from Figure 8.1 the within-age score distributions for the RPM (and, according to a personal communication from Robert Thorndike, the subscales of the Stanford-Binet test) are generally not Gaussian and are, indeed, often bimodal. Second, it does not encourage enquiry into whether there may be differential trends at different ability levels.

To avoid these (and other) problems, the normative data for the RPM have always been presented in the form of tables showing the raw scores required to do better than 5%, 10%, 25%, 50%, 75%, 90% and 95% of the population of a particular age. Table 8.1, giving the overall results from the 1979 British norming study among young people, is typical of the output.

One further word of explanation is appropriate. In order minimise the – usually fairly large – effects of sampling error, the figures in these tables have typically been smoothed by drawing a graph for each percentile curve which, as far as possible, (a) equalises the deviations of the raw scores above and below the line, (b) gives most weight to the most reliable data (the data for the oldest and youngest age groups in a sample is typically distorted by factors such as advancement and retention policies in education), and (c) reflects the trends in the most reliable data.

The need to smooth the data to minimise the effects of random variance and especially sampling error may be seen from a glance at Figures 8.5 and 8.9. The populations on which these graphs are based are not correctly described as "samples" at all because they consist of virtually the entire populations of successive age cohorts of Belgian men. Nevertheless, failure to graph the data would obscure the general trends that are so striking and there would always be the temptation to focus on explaining fluctuations around the general trend which, given the huge numbers, would always turn out to be overwhelmingly "statistically significant". When the data are based on smaller numbers and sampling error comes into play, the results become even more irregular (illustrative data will be found in Table B1 in Raven (2000b).

The importance, of, when drawing these graphs, giving more weight to the most reliable data can be reinforced by noting that, if one has 50 people in an age category, the 5^{th} and 95^{th} percentiles are the scores

lying half way between those obtained by the 2nd and 3rd person in the respective tails of the distribution. On their own, these estimates are therefore necessarily extremely unreliable.

The vitally important conclusions to be drawn from this digression are that the presentation of unsmoothed raw data can (i) lead to diversionary enquiries into the causes of chance fluctuations and (ii) when used as reference data against which to view the scores of individuals or experimental groups, to seriously misleading evaluations.

Table 8.1 (above) shows the (smoothed) raw scores corresponding to the 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, and 95^{th} percentiles for each age group in the 1979 British Standardisation.

When the 1979 U.K. data were broken down by a series of demographic variables (region, socio-economic status etc.) and analysed using multiple regression techniques (see Raven, 1981; 2000b) it emerged that 2.6 % of the SPM variance was accounted for by region. However, when the effect of Socio-Economic Status was partialled out, this dropped to 0.5%. Thus regional variation *per se* seems to be of little importance. SES on its own accounted for 4.8% of the variance. However, since age accounted for 46% of the variance, SES accounted for 8.9% of the variance which is not attributable to age. This is equivalent to a within-age correlation between Socio-Economic Status and the SPM of .30.

Population balance assessed via SES *is* therefore something that must be taken into account when comparing one set of results with another or when seeking to generalise from one population to another.

SPM score correlated .68 with age. Thus, more than half the variance was *not* "explained" by age. It is not, therefore, true that the tests simply measure "intellectual maturity."

As in the 1938 standardisation, Item Characteristic Curve (Item Response Theory or Rasch-type) based item analyses were carried out separately within each socioeconomic and age group. While the detailed figures take up too much space to present here, it may be noted (a) that the ICCs for individual items were remarkably similar to those published 40 years earlier and (b) that, as can be seen from the summary data presented in Table 8.2, the items scaled in much the same way for children from a variety of different backgrounds. . More recently, Vodegel-Matzen (1994) has shown that making the items more "realistic" (i.e. using hats, bananas etc. instead of abstract figures) while retaining the logic of the items makes the items easier for everyone, but changes neither the order of the items nor the order of individuals.

The conclusion is clear and vitally important: It is not possible to explain away differences in the mean scores of these groups on the grounds that, in any general sense, the test is "foreign to the way of thought of children from certain backgrounds." With certain important group and individual exceptions which will not be discussed here, the test generates orderly data which, on these grounds alone, must have some meaning. Differences between groups cannot be dismissed as "meaningless." They merit investigation and explanation.

U.S. Standardisations in the 1980s

Between 1983 and 1989 some 50 norming studies were carried out within school districts spread across the United States of America (J. Raven, 2000a). Within each district the sample was, as far as practicable, representative of the district. The specific sampling procedure employed varied from district to district, but, for reasons discussed in Appendix B, in no case were quota sampling procedures employed. (The sampling procedure adopted in each district is described in the previously mentioned publications.) Altogether more than 60,000 students aged 5 to 18 years were tested.

The norms which were obtained varied markedly from one school district to another and, within districts, between socioeconomic and ethnic groups.

As is illustrated in Tables 8.3 and 8.4 both ethnicity and socioeconomic status seemed to make independent contributions to the within-district variances. (It is important to note that the application of multiple regression

Table 8.2. *Standard Progressive Matrices*
1979 British Standardisation
Correlations Between Item Difficulties Calculated Separately for Young People from Different Socioeconomic Backgrounds
(Decimal points omitted and rounded to two decimal places)

SES	1 (High)	2	3	4	5	6	7	8 (Low)
1 (High)								
2	99							
3	99	99						
4	98	99	99					
5	97	98	99	99				
6	98	99	99	99	99			
7	95	96	98	98	99	98		
8 (Low)	95	96	98	98	99	99	99	

techniques is not strictly legitimate not only because the distributions are not Gaussian, but also because the independent variables are correlated with each other.) Accordingly, as shown in Table 8.4, the regressions were run twice with the independent variables entered in different orders: once with SES partialled out first and once with ethnicity partialled out first. Whichever way round the variables are entered, there is variance left to be explained by the other.

Differences between the norms for *school systems* catering for white students of differing socioeconomic status were as great as the ethnic differences within school districts.

The differences between the norms for school systems catering for different proportions of white, black and native American children seem to correspond to differences in published statistics on birth weight, infant mortality, and the incidence of serious childhood illness as published by the U.S. Bureau of the Census (United States Government, Bureau of the Census, 1984).

Within a number of school districts which had enough students of differing ethnicity to make the process legitimate, item analyses were run separately among different ethnic groups. One, fairly typical, example of the outcome is shown in Table 8.5. It follows from results like these (which duplicate those published by Jensen (1974) that the test works in the same way – measures the same thing – in each group. In addition, as illustrated in Figure 8.2, Hoffman (1983, 1990) demonstrated that the regression lines of RPM on various types of academic achievement for different ethnic groups were (to all intents and purposes) parallel – although having different intercepts. (Although the regression lines for mathematics shown in Figure 8.2 diverge while those for reading converge, these are only two examples. Overall, some diverge and some converge in such a way that it becomes clear that the general conclusion is that they are parallel.) Thus, while ethnic groups score at different levels on both achievement and matrix tests, the RPM has equal predictive validity within each group. Similar results were again reported by Jensen (1974).

Table 8.3. *Standard Progressive Matrices*
1986 Adolescent Percentile Norms for Ethnic Groups in Westown (U.S.) in the Context of 1979 British Data (Smoothed)

		Age in Years								
Percentile		12½	13	13½	14	14½	15	15½	16	16½
	UK	53	54	54	55	56	57	57	–	–
	Anglo	51	52	53	54	55	56	57	58	59
95	Asian	53	54	54	54	55	55	56	57	57
	Hisp	48	49	49	50	51	52	53	53	53
	Black	47	48	49	50	51	52	53	54	54
	UK	51	52	53	54	54	55	55	–	–
	Anglo	50	51	52	53	54	55	56	57	57
90	Asian	50	51	51	52	53	53	54	55	55
	Hisp	45	46	47	48	49	50	51	52	52
	Black	45	46	47	48	49	50	51	52	52
	UK	47	49	49	50	50	51	51	–	–
	Anglo	46	47	47	48	50	52	53	54	54
75	Asian	46	47	48	48	49	50	50	51	52
	Hisp	42	43	44	45	46	47	48	49	50
	Black	42	42	42	42	44	45	46	49	49
	UK	42	43	44	45	46	47	47	–	–
	Anglo	41	42	43	44	47	48	48	48	49
50	Asian	42	43	43	43	44	45	46	47	48
	Hisp	37	38	39	40	41	42	43	44	45
	Black	36	36	37	38	39	40	41	43	44
	UK	38	39	41	42	42	42	42	–	–
	Anglo	37	38	39	40	42	44	45	45	45
25	Asian	35	35	36	36	37	38	40	42	43
	Hisp	32	33	34	35	36	37	39	39	40
	Black	29	30	31	32	33	34	35	36	37
	UK	32	33	35	36	36	36	36	–	–
	Anglo	32	33	34	35	36	38	40	40	40
10	Asian	24	25	26	27	29	30	31	32	33
	Hisp	24	25	26	27	28	29	30	31	32
	Black	25	26	27	28	29	30	31	32	33
	UK	27	28	29	30	33	33	33	–	–
	Anglo	27	28	29	30	32	34	36	37	38
5	Asian	17	18	19	20	23	25	26	28	29
	Hisp	20	21	22	23	23	23	24	25	26
	Black	12	15	17	19	21	23	25	26	26
	UK	174	185	180	196	189	191	171	–	–
n	Anglo	46	59	44	52	53	36	56	40	49
(unweighted)	Asian	31	42	47	48	48	38	55	27	55
	Hisp	35	44	52	45	52	35	48	34	45
	Black	37	57	54	53	39	45	48	42	47

Note. The town name "Westown" was chosen, at the request of the school district, to preserve anonymity.

Studies with The Mill Hill Vocabulary Scale

So far, we have considered studies with the RPM. We turn now to the MHV.

The 1979 British Standardisation

The *Mill Hill Vocabulary Scale* (MHV) was standardised alongside the SPM in the 1979 study among young people, the sampling for which has already been described. As was the case with the SPM, there was no variance in MHV scores with region once the effect of Socio-Economic Status (SES) was partialled out (See Raven 1981; 2000b). SES explained 16.2% of the non-age-explained variance. MHV scores are, therefore, more related to background SES than SPM scores. Age accounted for 58% of the MHV variance. MHV scores did not plateau in adolescence in the same way as SPM scores; growth continued at approximately one and a half words per six-month interval through to age 15 ½ years.

As with the RPM, separate item analyses were carried out within eight SES groups. The reproducibility of the Scale properties across groups was again very high, averaging .97. The order in which children acquire knowledge of the meaning of words is therefore no more (and no less) affected by home background than is their ability to solve matrix problems. It would appear to be untrue that children from different backgrounds learn different subsets of dictionary words.

Table 8.4. *Standard Progressive Matrices*
US 1986 Data for Adolescents in Westown
Contributions of Ethnicity and Socio Economic Status to Total Variance

	Simple R	Mult. R	R sq.	R sq. change	Beta	Beta sq.
Age	29	29	8	8	27	7
Father's SES	-31	41	16	8	-20	4
Black	24	46	21	5	26	7
Hispanic	14	48	23	2	15	2
Asian	-04	48	23	0	0	0
Age	29	29	8	8	27	7
Black	24	38	14	6	26	7
Hispanic	14	44	19	5	15	2

Table 8.5. *Standard Progressive Matrices*
Correlations Between Item Difficulties Calculated Separately Within Specified Groups (Decimal point omitted and rounded to two decimal places)

		Westown Black	Westown White	Westown Hispanic	Westown Asian	Westown All	Des Moines	China
Westown	Black							
	White	98						
	Hispanic	100	98					
	Asian	98	99	98				
	All	99	99	100	99			
Des Moines		99	97	99	97	99		
China		95	94	94	96	95	96	
UK 1979		99	97	99	98	99	99	97

Standardisations in the US in the mid 80s

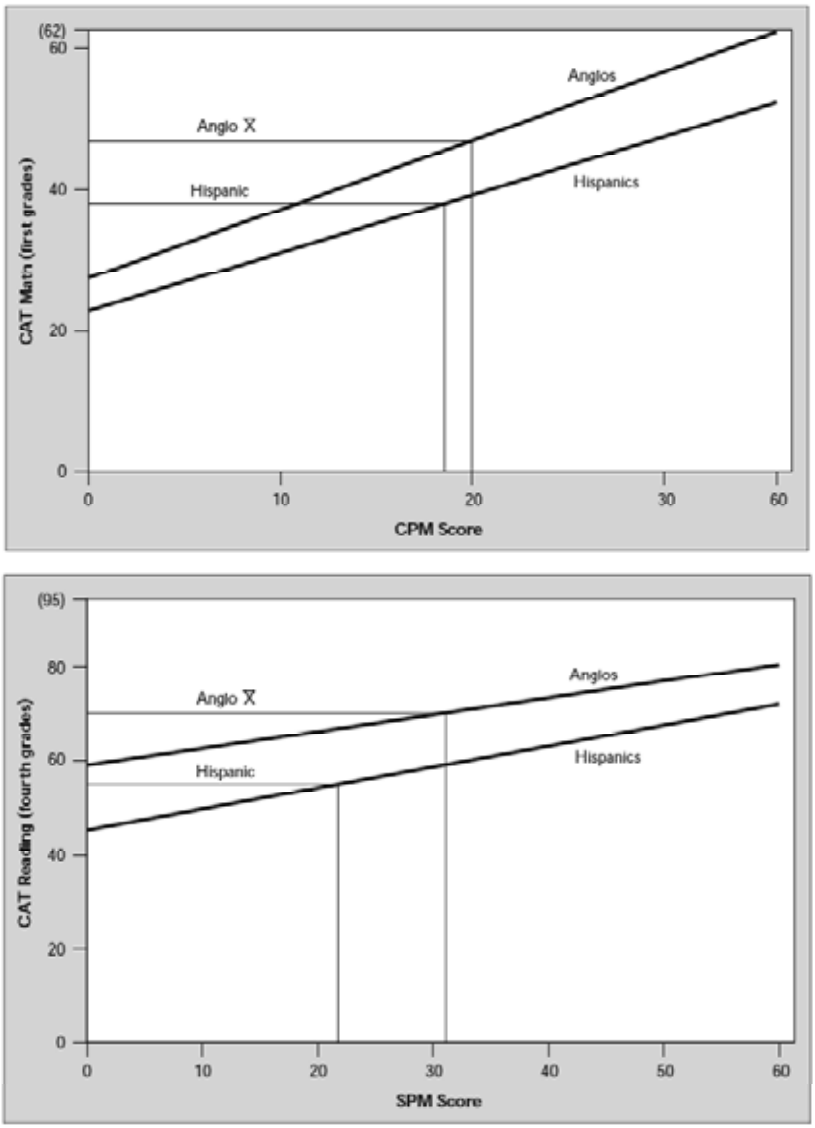
Many of the U.S. school districts that collected norms for the RPM did not administer the MHV. Nevertheless, as can be seen from Table 8.6, the overall U.S. norms for schoolchildren calculated from the data that were accumulated again lagged behind the international figures. However, the U.S. White norms once more corresponded fairly closely to those available for other cultures. As with the SPM, and as can be seen from Table 8.7, the test scaled in much the same way for (English speaking) students from different socioeconomic and ethnic backgrounds: Thus students from some backgrounds do *not* learn many of the kinds of word included in the Scale that are unknown to other cultural groups.

Studies of Adults

Standardisation in the U.K. in the Mid 1940s

The U.K. adult norms for the SPM that were published in the late 1940s and which formed the main reference data used worldwide for more than half a century were derived from a number of studies conducted between 1939 and 1947. Each of these samples and the way in which the resulting data were consolidated is described in some detail in Raven (2000b). Figure 8.3 shows the resulting norms in the form of a graph. Although the data were in fact collected over a number of years, they will, for convenience, hereinafter be referred to as the "1942 U.K. adult

Figure 8.2. *Coloured and Standard Progressive Matrices*
Sample regressions of the sub-tests of California Achievement Test on the RPM for Anglos and Hispanics in Douglas, Arizona.



Note: The upper figure shows the regressions of CAT Mathematics scores on the Coloured Progressive Matrices among first grade students. The lower figure shows the regressions of CAT Reading scores on the Standard Progressive Matrices among fourth grade students. (Redrawn from Hoffman, 1990.)

Table 8.6. *Mill Hill Vocabulary Scale*
1986 Adolescent Percentile Norms for the U.S.A. in the Context of 1979 British Data (Smoothed)

	Age in years (months)																			
	6½ 6(3) to 6(8)		7 6(9) to 7(2)		7½ 7(3) to 7(8)		8 7(9) to 8(2)		8½ 8(3) to 8(8)		9 8(9) to 9(2)		9½ 9(3) to 9(8)		10 9(9) to 10(2)		10½ 10(3) to 10(8)		11 10(9) to 11(2)	
Percentile	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**
95	23	**22**	24	**23**	26	**25**	28	**27**	30	**29**	32	**31**	34	**33**	36	**35**	38	**37**	40	**39**
90	21	**20**	22	**21**	24	**23**	26	**25**	28	**27**	30	**29**	32	**31**	34	**33**	35	**34**	37	**36**
75	18	**17**	19	**18**	20	**19**	22	**20**	24	**22**	26	**24**	28	**26**	30	**28**	32	**30**	34	**32**
50	**13**	**12**	**14**	**13**	**15**	**14**	**18**	**16**	**20**	**18**	**22**	**20**	**24**	**22**	**26**	**23**	**27**	**25**	**29**	**27**
25	8	**7**	9	**8**	10	**10**	13	**12**	15	**14**	17	**16**	19	**18**	20	**19**	22	**20**	24	**21**
10	6	**7**	6	**5**	7	**6**	9	**7**	10	**9**	11	**10**	13	**12**	15	**14**	18	**15**	20	**17**
5	5	**5**	5	**5**	5	**5**	6	**5**	7	**6**	7	**7**	9	**8**	11	**10**	13	**11**	15	**13**
n	103		135		149		175		157		183		205		179		209		200	
	11½ 11(3) to 11(8)		12 11(9) to 12(2)		12½ 12(3) to 12(8)		13 12(9) to 13(2)		13½ 13(3) to 13(8)		14 13(9) to 14(2)		14½ 14(3) to 14(8)		15 14(9) to 15(2)		15½ 15(3) to 15(8)		16 15(9) to 16(2)	16½ 16(3) to 16(8)
Percentile	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	US	**US**
95	42	**41**	44	**43**	45	**45**	46	**47**	48	**49**	52	**51**	54	**53**	56	**55**	57	**56**	57	**57**
90	39	**38**	41	**40**	43	**42**	45	**44**	47	**46**	50	**48**	52	**50**	53	**52**	54	**53**	53	**54**
75	35	**33**	37	**34**	38	**36**	40	**38**	42	**40**	44	**42**	46	**44**	47	**45**	49	**46**	47	**48**
50	**31**	**28**	**32**	**30**	**33**	**32**	**35**	**33**	**36**	**34**	**38**	**36**	**40**	**38**	**41**	**39**	**43**	**40**	**42**	**43**
25	25	**23**	27	**25**	28	**26**	30	**28**	32	**30**	34	**32**	36	**33**	36	**34**	37	**35**	36	**36**
10	21	**18**	22	**19**	23	**21**	25	**23**	27	**24**	29	**26**	30	**27**	31	**28**	32	**29**	30	**30**
5	16	**14**	17	**16**	19	**17**	21	**19**	24	**21**	26	**23**	28	**25**	28	**26**	29	**27**	28	**28**
n	167		173		179		192		195		201		198		197		185			

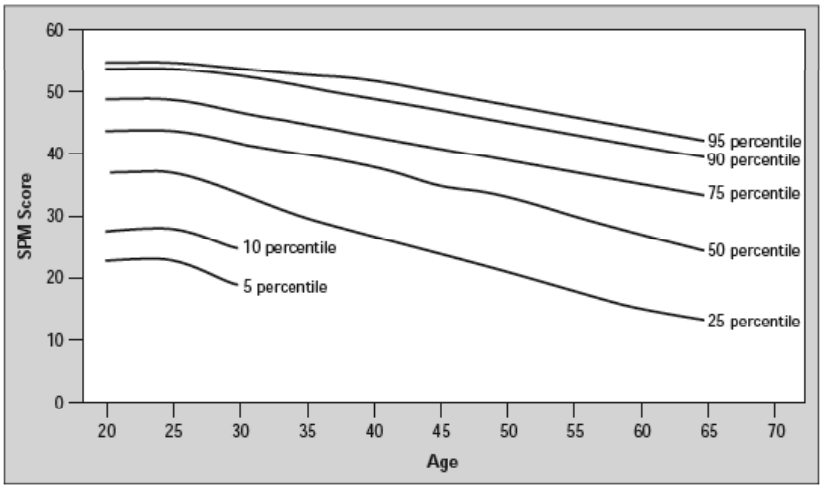
Note. US figures estimated on the basis of data available Summer 1986. The studies on which these norms are based are detailed in Raven et al. (1990). These show that the norms vary considerably between school districts, and, within districts, between ethnic groups.

norms". It will be seen that the scores obtained by successive age groups (after 20) become progressively lower. However, the scores of the less able appear to "decline" most. It must however, be stressed that the data for all age groups were, in essence, all collected at the same time. The graphs are *not* based on data collected from the same people as they got older. In other words they are based on cross-sectional as distinct from longitudinal data. As we shall see, interpreting the data *as if* they were longitudinal data has resulted in serious misunderstandings.

Table 8.7. *Mill Hill Vocabulary Scale*
Correlations Between Item Difficulties Calculated Separately Within Specified Groups (Decimal point omitted and rounded to two decimal places)

	Westown Black	Westown White	Westown Hispanic	Westown Asian
Westown Black				
Westown White	97			
Westown Hisp.	100	97		
Westown Asian	99	97	99	

Figure 8.3. *Standard Progressive Matrices*
The Apparent Decline in Scores as Age Increases as Documented in Typical Cross-Sectional Studies
UK Standardisation, circa 1942



Note: The graphs plot the cross-sectional norms for people of increasing age all tested, as described in Raven (2000a), in a series of studies conducted around 1942.

1992 Standardisation in Dumfries, Scotland

The first re-standardisation of the *Standard Progressive Matrices* in the U.K. (which was combined with the first ever standardisation of the *Advanced* Progressive Matrices on a general adult population) was carried out in Dumfries, Scotland, in 1992. Dumfries was chosen because: (a) the studies conducted among young people described earlier had made it clear that the most important variable to take into account in the choice of location was the balance of SES groups in the population, (b) these same studies had shown that the RPM norms obtained for the Borders region of Scotland – itself an area with a demographic structure which matches that of the U.K. as a whole – *did* correspond to those of the U.K. as a whole, and (c) the town of Dumfries recommended itself as a possible site for an adult standardisation because (i) much of the data collected with the RPM over the past 50 years (including two major standardisations of the *Coloured* Progressive Matrices) had been collected there and had stood up well in comparison with data collected elsewhere, (ii) it had itself a demographic structure which approximated that of the U.K., and (iii) it was geographically of such a size as to be easily traversed in search of named adults selected by systematic sampling procedures from a full list of adult residents. The procedures used for selecting, contacting, and testing the respondents, together with the response rates, are described in some detail in Raven (2000b).

1993 Standardisation in Des Moines, Iowa

Following the success of the Dumfries study, an exactly parallel study was carried out in Des Moines, Iowa. Des Moines is recognised as one of four U.S. cities having demographic compositions approximating the U.S. as a whole and is therefore widely used by researchers seeking a microcosm of America (*American Demographics,* May, 1985, pp. 38-42). While it is, of course, impossible for any one city – however closely its crude demographic statistics may correspond to the whole country – to match the whole at a detailed level, the studies conducted with the RPM among schoolchildren in the U.S. in the 1980s had in fact confirmed that the norms for Des Moines did approximate to those for the U.S. as a whole (J. Raven, 1989). Once again, the procedures used to draw the sample and contact potential respondents (together with response rates) are described in Raven (2000b).

Comparison of Dumfries and Des Moines Data

The norms for both the *Standard* and *Advanced* Progressive Matrices tests that were obtained in Dumfries and Des Moines were compared with each other. The APM norms for Des Moines are compared with those from Dumfries in Table 8.8. It can be seen that the Des Moines adult norms fall much where our previous research with the larger populations of schoolchildren would lead one to expect. The upper percentiles for Des Moines closely approximate those obtained in the U.K., while the 50th and lower percentiles – and especially the latter – lag behind, at least up to age 50. Nevertheless, our extensive research among young people (J. Raven, 2000a; J. Raven & Court, 1989) does suggest that the lower percentiles for the U.S. as a whole should lag further behind the U.K. norms than did those obtained in Des Moines. A number of possible explanations of this are discussed in J. Raven et al. (2000, updated 2004) but, whatever the explanation, the main point to be made here is that these Des Moines norms are probably above those which would have been obtained had a random sample of the entire U.S. population been tested.

Table 8.9 shows that the adult norms for the MHV for Dumfries and Des Moines were also similar.

The Effects of Date of Birth

It is time now to take another look at the variation in scores over time. To anticipate the outcome, we will find ourselves re-interpreting the data obtained from the previously mentioned cross-sectional studies which had been thought to show a significant decline in eductive ability, but less decline in reproductive ability, with increasing age. Our conclusion will be that a more accurate interpretation of these data is that most human abilities (including, for example, athletic ability), *but not reproductive ability* have improved rather dramatically over the past century and that this reveals a hitherto unsuspected effect of the environment on these scores. Yet the puzzle is not what it is usually taken to be. Since *most* abilities are improving, the question is why *reproductive* abilities show such little change despite investments in education and the mass media.

Eductive ability.

Figure 8.4 displays the 1979 normative data for the SPM which were derived from the previously described nationwide study of young people in

Table 8.8. *Advanced Progressive Matrices, Set II (Untimed)*
1993 Adult Percentile Norms for Des Moines, Iowa (U.S.) in the Context of 1992 Dumfries (U.K.) Data

	Age in years																					
	20		25		30		35		40		45		50		55		60		65		70	
	18		23		28		33		38		43		48		53		58		63		68+	
	to		to		to		to		to		to		to		to		to		to			
Perc-	22		27		32		37		42		47		52		57		62		67			
entile	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**
95	33	**32**	33	**32**	33	**32**	33	**32**	32	**32**	32	**32**	31	**31**	31	**30**	31	**29**	30	**27**	26	**25**
90	31	**30**	31	**30**	31	**30**	31	**30**	30	**30**	30	**30**	29	**29**	28	**28**	27	**27**	26	**25**	22	**23**
75	27	**27**	27	**27**	27	**27**	27	**26**	26	**26**	25	**26**	24	**26**	23	**25**	22	**24**	21	**22**	18	**19**
50	**22**	**20**	**22**	**20**	**22**	**20**	**21**	**19**	**20**	**19**	**19**	**19**	**18**	**19**	**17**	**18**	**16**	**16**	**15**	**14**	**13**	**12**
25	17	**15**	17	**15**	17	**15**	16	**15**	15	**15**	14	**14**	13	**14**	12	**13**	11	**12**	10	**10**	7	**8**
10	12	**10**	12	**10**	12	**10**	11	**10**	10	**10**	10	**10**	9	**9**	8	**8**	7	**7**	6	**6**	4	**4**
5	9	**7**	9	**7**	9	**7**	8	**7**	7	**7**	6	**7**	5	**6**	4	**5**	4	**4**	3	**3**	2	**1**
n	58	**28**	71	**53**	84	**72**	69	**77**	54	**121**	67	**69**	54	**33**	39	**36**	46	**27**	43	**33**	44	**54**

Note. Tests completed at leisure.

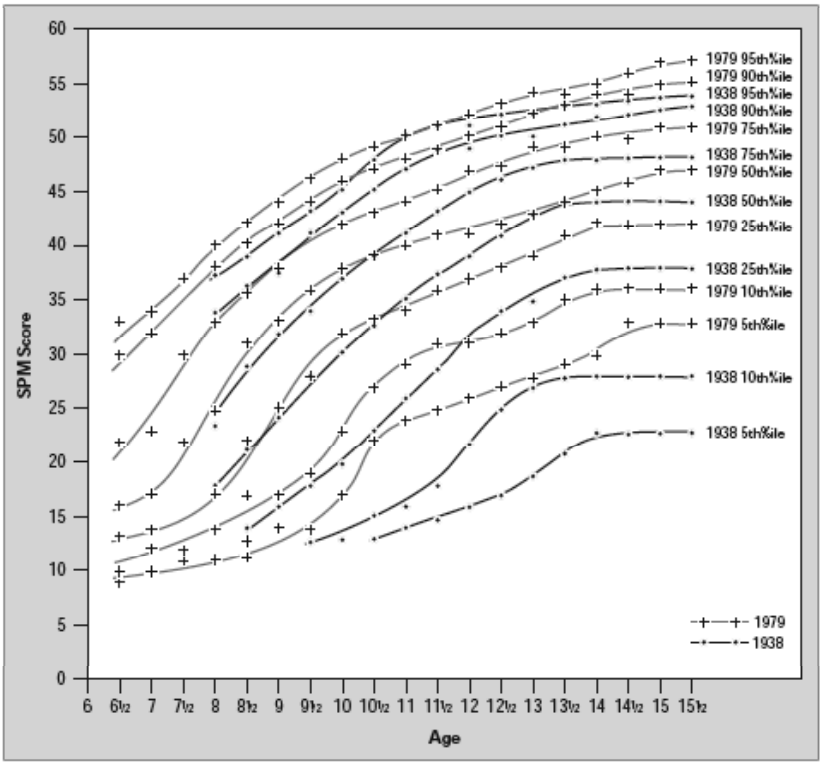
Table 8.9. *Mill Hill Vocabulary Scale, Forms 1 and 2 (Untimed)*
1993 Adult Percentile Norms for Des Moines, Iowa (U.S.) in the Context of 1992 Dumfries (U.K.) Data

	Age in Years																					
	20		25		30		35		40		45		50		55		60		65		70	
	18		23		28		33		38		43		48		53		58		63		68+	
	to		to		to		to		to		to		to		to		to		to			
Perc-	22		27		32		37		42		47		52		57		62		67			
	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**	UK	**US**
95	67	**69**	69	**70**	71	**71**	73	**72**	75	**73**	77	**74**	79	**75**	81	**76**	83	**77**	85	**78**	86	**77**
90	64	**63**	66	**65**	68	**66**	70	**67**	72	**68**	74	**69**	76	**70**	78	**72**	80	**73**	82	**73**	82	**72**
75	59	**56**	61	**57**	63	**58**	65	**59**	67	**61**	68	**63**	70	**65**	71	**66**	73	**68**	75	**68**	74	**67**
50	**53**	**51**	**55**	**52**	**57**	**53**	**58**	**54**	**60**	**55**	**61**	**57**	**62**	**58**	**63**	**60**	**64**	**62**	**65**	**62**	**63**	**61**
25	46	**44**	48	**46**	50	**47**	52	**48**	54	**50**	55	**51**	56	**52**	56	**53**	56	**53**	56	**53**	53	**52**
10	38	**36**	42	**38**	44	**40**	47	**42**	49	**44**	49	**46**	49	**46**	49	**46**	49	**46**	45	**46**	36	**45**
5	28	**23**	32	**25**	36	**27**	40	**31**	43	**43**	43	**35**	43	**37**	43	**38**	41	**38**	33	**38**	24	**36**
n	56	**26**	69	**53**	81	**70**	69	**75**	53	**118**	60	**68**	49	**31**	38	**35**	44	**29**	41	**32**	38	**56**

Note. Tests completed at leisure.

Great Britain (presented in Table 8.1) in the context of the data obtained in the 1938 Ipswich study. It is important to emphasise that the graphs within the chart do not show the scores obtained by the same young people as they grew older: They show the percentile scores obtained from a cross-section of young people of different ages who were tested in the same year. If one compares the graphs from the 1938 sample (i.e. the heavy lines) with those for the same percentiles for those tested in the 1979 sample (the light lines) it is clear that the level at which the scores plateau in adolescence has increased markedly and that young people

Figure 8.4. *Standard Progressive Matrices.*
Graphed percentile norms for young people in Great Britain in 1938 and 1979.



Note: The graphs show the score obtained by young people of different ages and levels of ability in these 2 years. If one compares the graph of the 1938 norms (i.e., the heavy lines) with those for the same percentile in 1979 (the light lines), it is clear that the level at which the scores plateau in adolescence has increased markedly and that young people get higher scores at earlier ages. (Thus, in the case of the 5th percentile, 10 1/2-year-olds in 1979 obtained similar scores to those obtained by 14-year-olds in 1938)

get higher scores at earlier ages. (Thus, in the case of the 5th percentile, 10 ½ year olds in 1979 obtained similar scores to those obtained by 14 year olds in 1938.)

Martinolli (1990) with the CPM and Spicher (1993) with the SPM have demonstrated similar changes in the norms over time in Fribourg, Switzerland. Case and her colleagues (see Raven et. al. 2000, updated 2004) have likewise documented similar changes for Argentina from 1964 to 2000).

What these results do not show is whether the increase has been continuous and incremental or whether it occurred at a particular time, such as during the second world war.

Bouvier's (1969) data, derived from testing conscripts to the Belgian army each year from 1958 to 1967, and reproduced in Figure 8.5, suggest that the increase was steady rather than associated with any particular developments.

The SPM results from the 1992 adult standardisation in Dumfries are shown, plotted by date of birth, by the dotted lines in Figure 8.6. The dashed lines re-plot the scores obtained in the 1940s study previously shown in Figure 8.3.

Examination of the points at which the two sets of graphs interface (i.e. among people born in 1922, where the earlier data are particularly strong) reveals that both the mean and spread of scores were very similar regardless of whether they were derived from the sample tested in the early 1940s (when they were roughly 20 years old) or from the sample tested in Dumfries in 1992 (when they were 70 years old). Instead of showing a decline in scores with advancing age (which is what the data behind each of these sets of graphs – and other similar data – had previously been thought to demonstrate), what the Figure clearly shows is a regular and continuous increase in the scores obtained by people born in different years, with the scores of the younger and more able respondents being the maximum obtainable on the test.

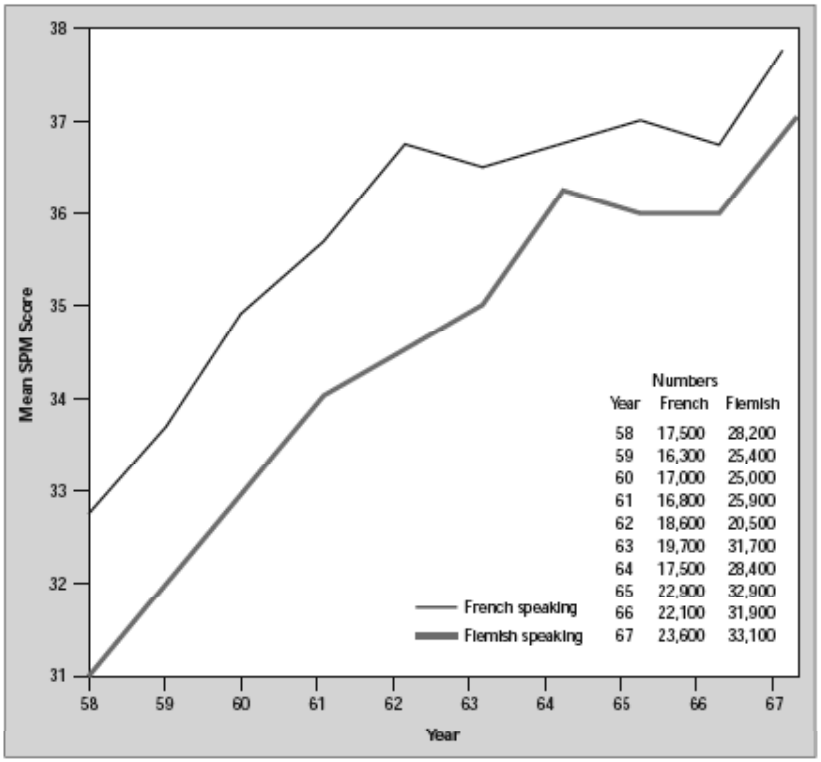
As previously noted, the continuity in the graphs derived from the two samples tested under different conditions in different places lends confidence to the adequacy of the data obtained in both studies.

This confidence is reinforced when data from a third – smaller – study conducted by Heron and Chown (1967) approximately halfway between the two studies already mentioned are introduced. The data have been superimposed on the graphs shown in Figure 8.6 in Figure 8.7. The graphs for the Heron and Chown data run straight through the point of interface between the graphs for the 1942 and 1992 data in the

previous Figure. They thus confirm the adequacy of the data from both the previously mentioned studies.

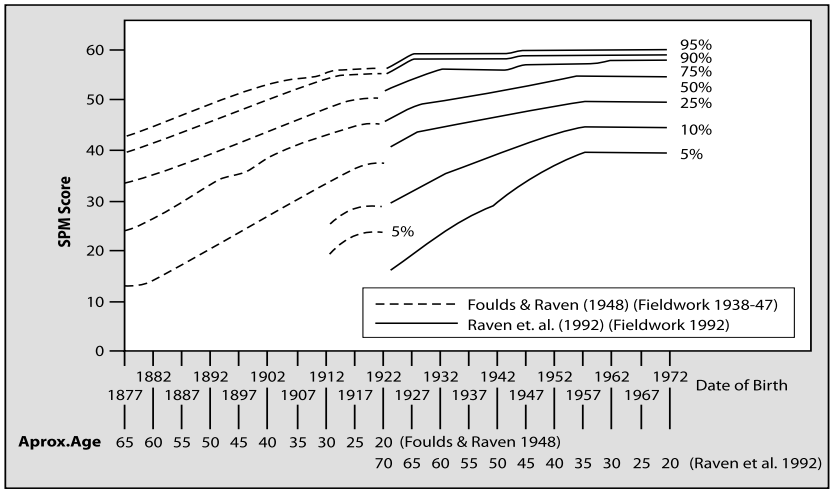
In an effort to guard against misleading conclusions being drawn from Figures 8.6 and 8.7, and because these Figures at first sight seem to confirm Teasdale and Owen's (1989) claim to "find no evidence of gains at the higher levels" it is important to note that the relatively small increase in scores among more able people born between 1922 and 1972 stems entirely from a ceiling effect on the SPM, which has only 60 items. As already mentioned, in both the Dumfries and Des Moines standardisations, the *Advanced* Progressive Matrices was standardised alongside the SPM. The APM norms for Dumfries for 1992 are compared with the 1962 norms for same test in Table 8.10. It is immediately

Figure 8.5. *Standard Progressive Matrices*
Mean scores of Belgian military conscripts from 1958 to 1967.



Note: The mean scores of French and Flemish-speaking recruits are graphed separately. (Redrawn from Bouvier, 1969, and reprinted with permission.)

Figure 8.6. *Standard Progressive Matrices*
100 years of eductive ability in Great Britain.
Graphed Percentile norms from the 1942 and 1992 standardisations plotted by date of birth.



Note: The Figure graphs the percentile norms obtained by adults of different ages (and thus dates of birth) on the Standard Progressive Matrices when a sample was tested circa 1942 (see comment on Figure 8.3) in one case and in 1992 in the other. The approximate age of people born in different years in the two samples is shown below.

Figure 8.7. *Standard Progressive Matrices*
UK standardisations, circa 1942, 1962 and 1992
100 years of eductive ability, including Heron & Chown's data from 1962.

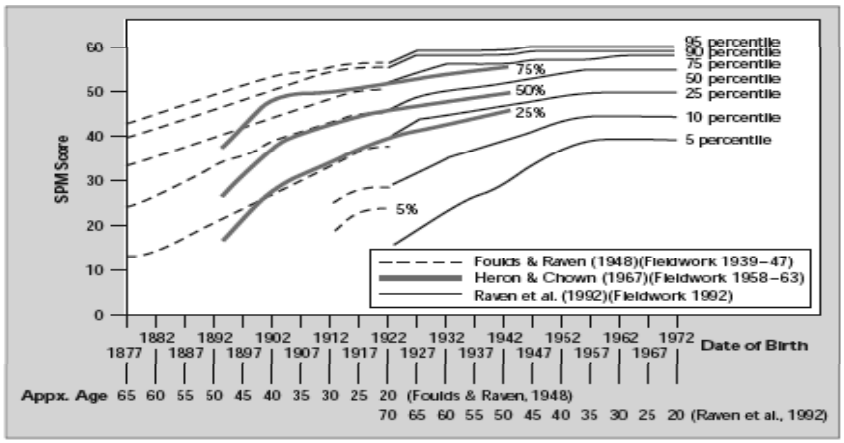


Table 8.10. *Advanced Progressive Matrices, Set II*
Comparison of 1992 and 1962 British Adult Percentile Norms

	Age in years					
	20		30		40	
Percentile	1962	**1992**	1962	**1992**	1962	**1992**
95	24	**33**	23	**33**	21	**32**
90	21	**31**	20	**31**	17	**30**
75	14	**27**	12	**27**	9	**26**
50	**9**	**22**	**7**	**22**	–	**20**

Note. The 1962 data (previously published in J. C. Raven, 1965) were estimated from the work of Foulds and Forbes, which was also published in J. C. Raven (1965). Since the test has 36 items and 8 options per item, scores of 6 or less verge on the chance level. There was therefore no point in publishing the lower percentiles in 1965.

obvious that the increase in scores evident in the lower percentiles in Figure 8.6 *has* been accompanied by major gains among the more able. (The enormous methodological difficulties which inhere in any attempt to isolate the relative size of gains at different points even on scales which conform to the Rasch model are discussed in the chapter by Prieler and Raven in this volume.) The effect was so great that the APM, which was originally developed to discriminate among the top 20% of the population, now offers an almost perfect Gaussian distribution across the entire adult population. Just as the entire distributions of height and athletic ability have moved up (with admittedly some change in shape), so has the entire distribution of eductive ability.

Reproductive ability

Turning now to reproductive ability, Table 8.11 compares the 1979 U.K. norms for adolescents on the *Mill Hill Vocabulary Scale* (MHV) with those obtained using the written test in Colchester in 1943. The 95th percentile has unmistakably dropped from 1943 to 1979. So has the 90th. The 75th has dropped, but the drop is less marked. The 50th is, to all intents and purposes, unchanged. The 25th had gone up. And the 10th and 5th show a marked increase.

Unfortunately, these apparently unambiguous results are not entirely confirmed when a comparison is made between the 1979 results and those obtained by oral administration of the MHV in 1943. Perhaps most importantly, whereas the comparison of the results obtained with the written test suggest a reduced variance in 1979, the comparison between the written test in 1979 and the oral test in 1943 indicate *increased*

Table 8.11. *Mill Hill Vocabulary Scale: Forms 1 and 2 (Self-Completed in Writing)* **UK data Norms for Adolescents from the 1979 Standardisation in the Context of 1943 Colchester Data**

	Age													
	11½		12		12½		13		13½		14		15	
Percentile	43	**79**	43	**79**	43	**79**	43	**79**	43	**79**	43	**79**	43	**79**
95	41	**42**	47	**44**	50	**45**	52	**46**	54	**48**	57	**52**	60	**56**
90	40	**39**	43	**41**	47	**43**	49	**45**	51	**47**	53	**50**	55	**53**
75	34	**35**	36	**37**	40	**38**	43	**40**	44	**42**	45	**44**	48	**47**
50	**29**	**31**	**31**	**32**	**33**	**33**	**35**	**35**	**37**	**36**	**38**	**38**	**40**	**41**
25	24	**25**	26	**27**	27	**28**	29	**30**	30	**32**	31	**34**	33	**36**
10	17	**21**	19	**22**	21	**23**	22	**25**	24	**27**	25	**29**	26	**31**
5	12	**16**	14	**17**	16	**19**	17	**21**	18	**24**	19	**26**	20	**28**

Note. Based on samples of 1,419 (1943 data) and 1,304 (1979 data).

variance, with more able pupils appearing to know still more and less able pupils knowing still less!

Figure 8.8 presents the U.K. adult data. The graphs plot the percentile *Mill Hill Vocabulary Scale* scores achieved by a cross-section of adults tested during the 1940s alongside those obtained by a cross section of the population in 1992. It will be seen that, although there appear to have been some changes – with the scores obtained by less able middle-aged adults seeming to have gone up most – the changes are nothing like as great as those which have occurred in eductive ability.

Bouvier's (1969) data (Figure 8.9) likewise reveal little change in vocabulary test scores over the period of his study, especially among the French speaking group.

All these results suggest that reproductive ability – at least as assessed from people's knowledge of words – has changed much less than might have been expected, and certainly a great deal less than eductive ability, over the period for which data are available.

Schaie (1983, 1994) and Thorndike (1977) likewise concluded that it is the reasoning components of "intelligence" which have been increasing most rapidly and consistently. Their data are particularly interesting in that they show that this is true whether "reasoning ability" is measured by verbal or nonverbal tests and whether reproductive ability is measured by vocabulary or other routine skills like word fluency. On the other hand, their data do suggest that knowledge of vocabulary has increased rather more than the above data would lead one to expect and that scores on tests which require these two abilities to different

Figure 8.8. *Stability and Change in Reproductive Ability Over Time*
Mill Hill Vocabulary Scale: Forms 1 and II
Graphed Norms from Cross Sectional Norming Studies
Conducted among British Adults circa 1945 and 1992

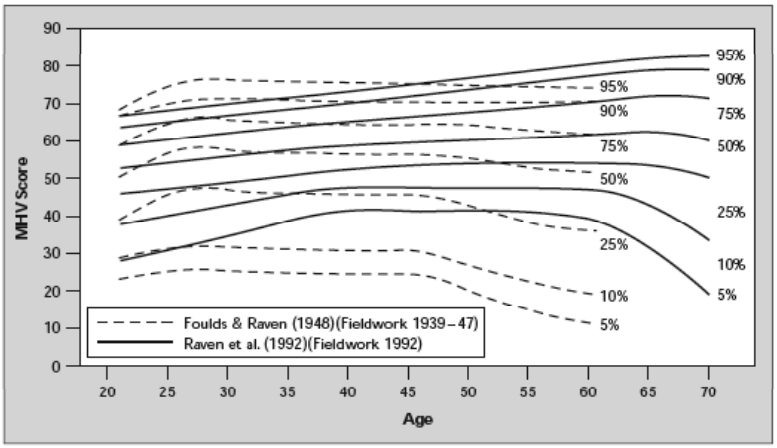
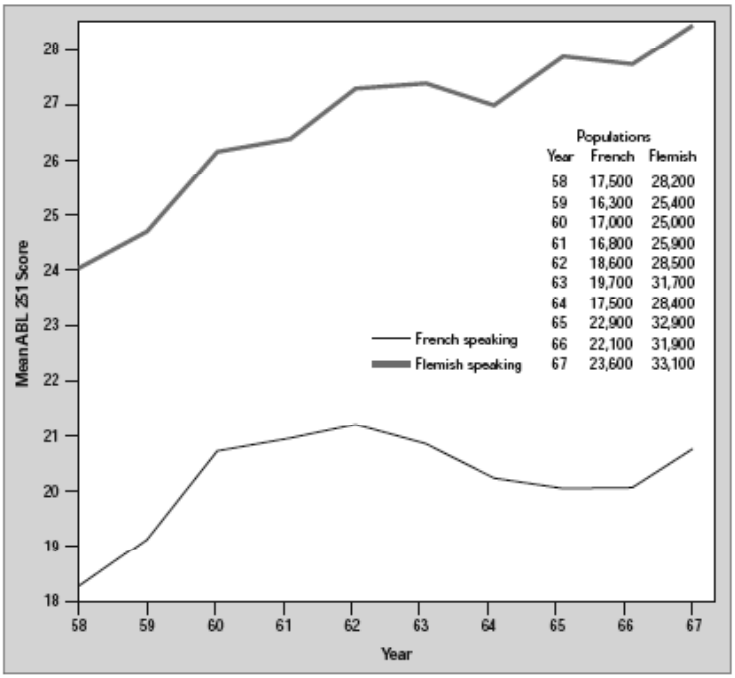


Figure 8.9. *Belgian Army Vocabulary Test (ABL 251)*
Mean scores of conscripts to the Belgian Army from 1958 to 1967.



Note: The mean scores of French and Flemish-speaking recruits are graphed separately. (Redrawn from Bouvier, 1969, and reprinted with permission.)

extents have increased in proportion to the extent to which they involve eductive ability. This inference has been strongly confirmed in Flynn's (2000) studies of the sub-scales of the WISC. The scores on these sub-tests have gone up in direct proportion to their loadings on an eductive ability factor. Schaie's position on these issues seems to have shifted over the years. In one publication (Schaie, 1983), he showed reasoning ability, whether measured by verbal or nonverbal tests increasing most, and numeric ability increasing and then declining, with other abilities falling in between. In a more recent article (Schaie, 1994), he presents graphs for what he calls "cohort gradients" for "latent abilities". According to these data, mean scores on "inductive reasoning" and "verbal memory" increased most steeply over the years. "Numeric ability" at first improved and then declined. His graph for what he calls "verbal ability" behaves somewhat similarly, but shows a later peak.

Discussion

It would appear from the results summarised above that there has been, and still is, considerable – if far from perfect – similarity in the SPM norms obtained in different societies with a tradition of literacy at any given point in time. However, in common with the scores on other tests, and especially those measuring eductive ability through verbal or nonverbal items – see for example Bouvier (1969), Thorndike (1975, 1977), Garfinkel and Thorndike (1976), and the large number of published and unpublished studies brought together by Flynn (1984, 1987) – there has been a continuous increase in the scores at all levels of ability over time.

The data on changes over time and the differences between ethnic groups naturally raise the question of what is responsible for these changes and differences. No one study – let alone any study of a correlational nature – can give a definite answer. But, since they do seem to make some hypotheses less likely and others more likely, it is worth summarising some of the data which bear on the question.

In what follows, the causes of the changes over time and the differences between ethnic groups will be considered simultaneously. On the one hand, the absence of cross-cultural differences in RPM scores between cultures which do differ on a variable which has been put forward as possible explanations of the changes over time make those explanations of the time differences less likely. On the other hand, variation in scores

between cultural groups which do differ on a variable which has also changed over time and been suggested as a possible explanation of the time differences strengthens the possibility of that variable playing a significant role in the process.

Thorndike (1977) and Garfinkel and Thorndike (1976) listed a number of possible explanations of the time trends. However, the data available on the *Progressive Matrices* do not really support any them. Thorndike suggested for example, that the acceleration in development may be due to earlier maturity. However, if maturity is a factor, the curves plotting the age norms for boys and girls separately should differ more than the data published in J. Raven (1981) and J. Raven et al. (2000a) shows that they do. These data show that, with the exception of an unexplained divergence between the two curves at age 11 (when there is a school change) the curves are virtually identical. Furthermore that divergence itself has not been confirmed when we have plotted similar graphs for, e.g., a range of U.S. school districts. Likewise, he suggested that the increases may have been due to changes in the nature of early school education, but the fact that there was little difference between the RPM norms obtained in Scotland and England in the 1979 standardisation suggests that this is unlikely – because Scottish infant education remains very formal (HMI, 1980). The minor difference between the Chinese and British norms likewise tends to disconfirm this contention. Indeed, some of the school systems for which norms are available do not admit children until they are eight years old, and, as Thorndike himself noted, the largest increases seem to have occurred among children of preschool age. Thorndike suggests that television may have had an effect. However television was widely available in Ireland when what can now be seen to be low Irish norms were collected. Greenfield's (1998) argument that the change is due to familiarity with icons and computer games likewise does not hold up because, as Schaie (1983) has shown, there has been a *huge* increase in scores on *verbal* measures of "reasoning" (or eductive) ability.

Others have suggested that the increases in RPM scores over time may be attributed to schools using matrix-type problems to teach "problem solving". However, Thorndike showed that performance on *all* the subscales of the Stanford-Binet had improved and that the greatest increases were among very young children who had not yet started school. In our own data there is little difference between the norms from cultures which differ markedly in the age at which children start school.

Flynn (having, in 1984, queried Thorndike's hypotheses concerning the Binet results) likewise concluded in his 1987 article that most of the common and obvious explanations of the RPM increase do not hold up. Among other things he showed, through a detailed analysis of de Leeuw and Meester's (1984) data, that changes in the amount of education people have could account for only one of the 20 IQ-point gain in RPM scores documented among servicemen. Changes in the intellectual quality of the home environment – at least insofar as it is indexed by SES – could account for little more.

In summary, then, most of the common explanations of the changes over time do not hold up: Where there is variation between cultures in a variable which potentially help to explain the change over time it is not accompanied by differences in RPM scores. Having, in this way, made such explanations *less* likely (although not, of course, ruling them out) it behoves us to look elsewhere.

A potentially more fruitful line of enquiry is suggested by the fact that the variation in mean scores between ethnic groups within the U.S. does seem to correspond to variation between the same groups in height, birth weight, and infant mortality. Height and birth weight have, like intelligence test scores, increased over the past 80 years (Floud, Wachter, & Gregory, 1990). These observations led us to suspect that the increase in RPM scores over time might be attributable to the same factors as have been responsible for increases in height and birth weight and for decline in infant mortality – that is, to improved nutrition, welfare, and hygiene. This statement does not, unfortunately, get us very far since what it is about the causes of these changes remains obscure. But then, so do the ways in which most drugs produce their effects: The fact that one cannot offer a complete explanation does not undermine the value of what one *can* do to clarify the position.

On looking for literature which bore on the nutritional hypothesis we unearthed a remarkable, although largely unpublished, study carried out in Aberdeen, Scotland, which showed that dietary and hygiene variables *do* influence RPM scores as well as birth weight and height (Baird & Scott, 1953; Scott, Illsley, & Thomson, 1956). In this study, calcium intake was used as an index of quality of diet. This had a marked impact on all three of the outcomes mentioned – and the relationship held up both within and between socioeconomic groups. Vernon (1969) had come to a similar conclusion.

More recently, Benton and Roberts (1988), Benton and Butts (1990), Benton and Cook (1991), Nidich, Morehead, Nidich, Sands,

and Sharma (1993), Schoenthaler, Amos, Doraz, Kelly, and Wakefield (1991), Schoenthaler, Amos, Eysenck, Peritz, and Yudkin (1991), and others have shown that vitamin and mineral supplements have a rapid and marked effect on some people's eductive, but not reproductive, ability scores.

Confirmation of the nutrition/hygiene hypothesis comes from the reviews by Pollitt and Saco-Pollitt (1996) and Sigman and Whaley (1998) of physical factors affecting intelligence. Low levels of iodine in diet had a major effect in the U.S. well into this century, and still do in other countries. Intestinal parasites have similar effects. Low oxygen pressures arising from high altitude also have a marked effect (cf. the previously mentioned Peruvian mountain norms).

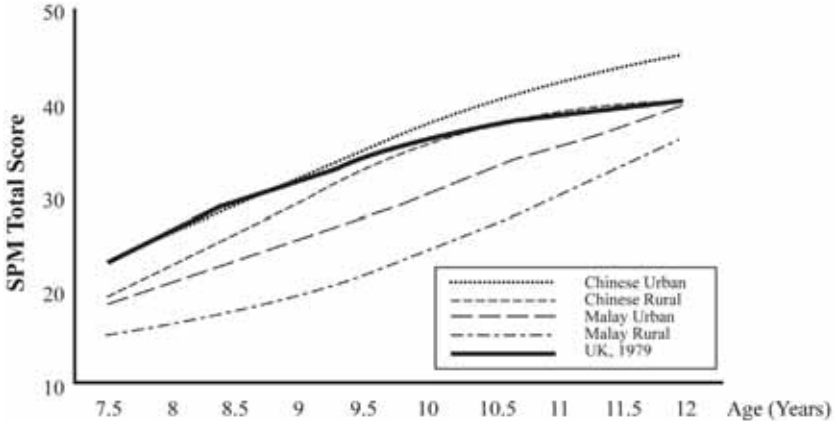
Although all these studies seem to support the contention that the differences over time and between cultural groups reflect the balance of vitamins available in the diet (*not*, pace Martorell, 1998, the absolute quantity of food) with isolated mountainous areas – which are unlikely to have extensive trade in agricultural products – having the lowest scores, this conclusion does not receive unequivocal support in the literature.

Chiam's data is particularly disconcerting. In two senses. First, in order to determine what they had been eating, Chiam (1995) collected samples of faeces from a cross-section of children for whom a variety of test scores (but not RPM scores) were available. She found that diet was unrelated to any of them. In another study (Chiam, 1994) of 5,412 children aged 7 ½ to 12 years in contrasting areas of Malaysia, she found that, as Figure 8.10 shows, the RPM norms for the Chinese in Malaysia corresponded to the international norms while those for the Malay population did not. If diet were responsible for the differences, one would expect them to be reduced when such things as the effects of urban and rural residence and SES were partialled out. Yet, although, as can be seen, there were differences between the urban and rural norms for both groups, this variable did not account for the differences between the groups. Neither difference was accounted for by education or socioeconomic status. Once again, then, the analyses render less likely the most commonly offered explanations of group differences and changes over time but in this case also throw doubt on the dietary hypothesis.

Some implications

It seems from this review of available literature that most of the most commonly offered explanations of the changes over time (including those promoted by such authors as Greenfield & Williams in Neisser's 1998,

Figure 8.10. *Standard Progressive Matrices*
Mean Scores For Young People in Malaysia by Ethnicity and Area of Residence
1992 Data



From Chiam (1994)

book) are unlikely to be true. Even the most likely hypothesis – the diet/welfare/hygiene hypothesis – does not receive unequivocal support. But whatever their *causes*, the changes which have been documented point to a clear, major, and previously unsuspected effect *of the environment* on eductive ability.

It follows from this that, despite their persistence over time as the scores of *all* groups have been going up (see also the persistence of the French-Flemish speaking difference in Belgium shown in Figures 8.5 and 8.9), the differences between ethnic groups within the U.S. cannot be regarded as immutable. It was, of course, precisely *this* point – and not some more general argument about the nature of "intelligence" – that Flynn sought to establish when he set about documenting and publicising the mutability of "intelligence" test scores. The point may be underlined by noting that most of the RPM norms for ethnic norms within the U.S. currently lie between the 1938 and 1979 British norms, that is, within the range over which environmentally induced change might not only have been theoretically possible but has actually been demonstrated.

Other features of the environment which make a difference

Although the effects are insufficient to explain the gross time and cultural differences discussed in this article, and although it would not be appropriate to present a thorough review of the relevant literature here,

there have been a number of empirical studies of factors which increase or decrease RPM scores and it is worth mentioning some of them as a counterpoint to simplistic hereditatrian and dysgenic arguments about "intelligence". The results surprise many psychologists. Eductive ability has turned out to be more easily influenced by appropriate educational and developmental experience than reproductive ability. However, the variables which influence the development of eductive ability are *not* the obvious cultural and socioeconomic variables which divide society and on which sociologists have focused so much attention. Acquired information *is* more influenced by these variables than is the ability to perceive and think clearly – but these background variables still account for only a small proportion of the total variance.

Eductive Ability, Upbringing, and Education

Many studies (e.g. Chan, 1981; McGillicuddy-DeLisi, 1985; McGillicuddy-DeLisi, DeLisi, Flaugher, & Sigel, 1987; J. Raven, 1980; Sigel & Kelley, 1988) have shown that the development of children's eductive ability is promoted if their parents involve them in their own thought processes. Such parents involve their children in their own attempts to make sense of difficult situations, as they use their feelings as a basis for "experimental" action, as they resolve value conflicts, and as they consider the long-term social consequences of their actions. All this necessitates that parents share with their children their own understanding of the workings of society and their role in it. The children are thereby presented with a thought process which is fundamentally conceptual, yet which also relates thought to action. Such parents are also more likely to treat their children with respect, and realise the need to earn (rather than demand) their children's respect. This leads them to initiate a cyclical process in which they discover just how competent their children really are and, as a result, become more willing to place them in situations which call for high-level competencies. The result is that their children have many opportunities to practice and develop these competencies. Such parents are more inclined to read to their children stories which bear on moral problems. The outcome is that the children empathise with the various characters in the books and are able to reach their own moral position. The importance of reading *to* children in the development of their moral character and analogical reasoning has been underlined in the work of Jackson (1986) and Vitz (1990).

J. Raven (1980, 1987, 1989) and Vygotsky (1978, 1981) have shown that the above is only part of a wider process whereby parents

who effectively nurture high-level competencies in their children tailor environments to the motives, incipient talents, and problems of their children. This is one way in which, as Plomin (1989) and Plomin and Daniels (1987) have shown, the within-family variance in children's environments becomes considerable and linked to variance in inherited characteristics in a way which markedly affects their development. As Scarr, Webber, Weinberg, and Wittig (1981) have noted, a similar effect is produced as children select themselves into different environments.

It follows from these observations that, if we wish to identify the genetic and environmental variables which influence psychological development, we will need to develop a more sophisticated model of the process.

The development of eductive ability in schools (but only in some cases measured by the RPM) has been studied by a number of researchers. Nickerson, Perkins, and Smith (1985) and Stallings and Kaskowitz (1974) found that the development of eductive ability is promoted by at least some forms of "open" or "progressive" education. Miller, Kohn, and Schooler (1985, 1986) and J. Raven, Johnstone, and Varley (1985) found that educational self-direction (i.e. pupils taking responsibility for their own education and moral decisions) and the undertaking of more complex educational activity (e.g. project-based, enquiry-oriented work) gave rise to a cyclical development in cognitive ability. Greater emphasis on self-direction and the development of new understanding fosters student competence, which in turn increases students' desire to gain more control over their destinies and encourages teachers' willingness to rely on their pupils' abilities.

Schooler, Mulatu, and Mesfin (2001), in the course of a 30-year follow up of a sample originally interviewed and tested in 1964, have confirmed their earlier longitudinal work (conducted mainly with Kohn) showing that substantively complex work improves intellectual functioning and, in a remarkable experimental study, Lovaglia, Lucas, Houser, Thye, and Markovsky (1998) have shown that even relatively minor, experimentally-induced, changes in perceived status produce significant (half-standard deviation) changes in RPM scores. (It may be worth noting that a change of this magnitude is greater than is typically achieved by training in the methods required to solve the problems.)

Having reviewed material demonstrating the importance of certain child development and educational practices in promoting the development of eductive ability, it is important to repeat that none of the psychological

and educational processes mentioned above produce effects sufficient to account for the inter-generational increase in RPM scores. Furthermore, none of the activities described in the studies published to date significantly reduce the variance within socioeconomic groups and within families. Yet the within-family variance amounts to two thirds of the variance in test scores. It therefore seems that the environmental factors which have most influence on eductive ability are not the psychological and educational variables with which psychologists have been preoccupied in the past, and they appear to have little effect on its heritability.

The Quest for Single-Factor Explanations

In the *General Introduction* to this book, we argued that the evidence that both the RPM and MHV conformed to the requirements of Item Response Theory showed that *eductive* and *reproductive* abilities are every bit as "real" and measurable as the "hardness" of substances in geology, "height" in physiology, "high jumping ability" in athletics, and "life expectancy" in actuarial accounting.

In the same chapter we included a graph showing that life expectancy for men in the U.K. has almost doubled over the century for which RPM data have been available. We suggested that this graph undermined most of the arguments Flynn had put forward in his attempt to undermine the meaningfulness of psychological tests: He had argued, for example, that back projection of the increases in RPM and other "intelligence" test scores to the time of the Greeks would mean that they must have been incredibly stupid and, since this could not have been the case, that the tests must be meaningless. Unfortunately for Flynn's case, the same arguments would apply to height, athletic ability, and life expectancy. (As will be seen in later chapters, we do not dispute his argument that the importance of these psychological abilities as determinants of life performance has been exaggerated.)

We used the same analogies to draw attention to the illogicality in other arguments which crop in discussions of the nature of eductive and reproductive abilities and the environmental variables which affect them. No one would use the scaleability of any of the previously mentioned variables as a basis from which to argue that the observable and measured variance must be due to a single underlying cause in the way in which many psychologists have argued that, because eductive ability is scaleable, the variance must be due to a single underlying factor like "speed of neural transmission". Nor would they be tempted to argue that, because

high jumping ability is clearly trainable, the measure is meaningless. Still less would they be inclined to set out to search for a single cause for the increases in height, high jumping ability, and life expectancy that have occurred over the years.

Yet there is one final, very intriguing, and almost completely overlooked question which emerges from this research. This is: Why is it that, when everything else has been going up … height, athletic ability, life expectancy … *reproductive* ability, whether measured by the MHV or any of a host of other measures that exist in this area, has been increasing hardly at all? This despite the massive investments that have been made in education, mass media, and information technology. In a sense, Flynn has focused the attention of psychologists on exactly the wrong question.

Concluding Asides

In concluding, it seems appropriate to draw attention to the seriousness of the errors which stem from the use of outdated norms. In the first place, it is obvious from Figure 8.6 that a score that would place a 50 year old tested in 1942 at the 95th percentile if judged against the 1942 norms would result in classification as at the 25th percentile if judged against today's norms. Such huge discrepancies in the interpretation of scores mean that the use of out-of-date norms cannot be justified: They are bad for the individuals concerned, bad for the organisations for which they work, and bad for society.

Still more serious, however, are the errors which arise from the adoption of out-of-date norms in research. The effectiveness of such things as educational enrichment programs is typically evaluated by comparing the scores obtained by experimental groups with published norms. When these norms are out-of-date, such experimental programs can only appear to be much more effective than they are.

References

Adams, E. A. (1952). *Analysis of Raven's Matrices Scores*: *Preliminary Report.* Surrey, England: Surrey Educational Research Association.

Angelini, A. L., Alves, I. C. B., Custodio, E. M., & Duarte, W. F. (1988). *Manual Matrizes Progressivas Coloridas.* Sao Paulo, Brazil: Casa do Psicologo.

Baird, D., & Scott, E. M. (1953). Intelligence and childbearing. *Eugenics Review, 45(3),* 139-154.

Benton, D., & Butts, J. (1990). Vitamin/mineral supplementation and intelligence. *The Lancet, 335*, 1158-1160.

Benton, D., & Cook, R. (1991). Vitamin and mineral supplements improve the intelligence scores and concentration of six-year-old children. *Personality and Individual Differences, 12*, 1151-1158.

Benton, D., & Roberts, G. (1988). Effect of vitamin and mineral supplementation on intelligence of a sample of schoolchildren. *The Lancet, 23*, January, 140-143.

Bouvier, U. (1969). *Evolution des Cotes a Quelques Test.* Belgium: Centre de Recherches, Forces Armees Belges.

Byrt, E., & Gill, P. E. (1973). *Standardisation of Raven's Standard Progressive Matrices and Mill Hill Vocabulary for the Irish Population*: *Ages 6-12.* Unpublished Master's Thesis, National University of Ireland, University College Cork.

Chan, J. (1981). Correlates of parent-child interaction and certain psychological variables among adolescents in Hong Kong. In J. L. M. Binnie-Dawson (Ed.), *Perspectives in Asian Cross-Cultural Psychology* (pp. 117-131). Lisle, Netherlands: Swets and Zeitlinger.

Chan, J. (1989). The use of Raven's Progressive Matrices in Hong Kong: A critical review. *Psychological Test Bulletin,* November, *2*(2), 40-45. Hawthorn, Victoria: ACER.

Chiam, H. K. (1994). *Is the Raven Progressive Matrices valid for Malaysians?* Paper presented to the 23rd International Congress of Applied Psychology, Madrid.

Chiam, H. K. (1995) *The standardisation of several tests in Malaysia.* Unpublished manuscript, School of Education, University of Malaya, Kuala Lumpur.

Court, J., & Raven, C. J. (2001). *A Researcher's Bibliography for Raven's Progressive Matrices and Mill Hill Vocabulary Scales.* Obtainable in hard copy and disk format from Susan Middleton, Harcourt Assessment, 19500 Bulverde Rd., San Antonio, Texas 78259, USA. <susan_middleton@harcourt.com>

Court, J. H., & Raven, J. (1995). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 7*: *Research and References*: *Summaries of Normative, Reliability, and Validity Studies and References to all Sections.* San Antonio, TX: Harcourt Assessment.

Dahlstrom, W. G. (1993). Tests: Small samples, large consequences. *American Psychologist, 48(4),* 393-399.

de Leeuw, J., & Meester, A. C. (1984). Over het intelligente-nderzoek bij de militaire keuringen vanaf 1925 tot henden. [Intelligence-as tested at selections for the military service from 1925 to the present]. *Mens en Maatschappij, 59,* 5-26.

de Lemos, M. M. (1984). A note on the Australian norms for the Standard Progressive Matrices. *Bulletin for Psychologists, (ACER Melbourne), 36,* 9-12.

de Lemos, M. M. (1989). The Australian re-standardisation of the Standard Progressive Matrices. *Psychological Test Bulletin,* November, *2*(2), 17-24. Hawthorn, Victoria: ACER.

Ferjencik, J. (1985). *Manual*: *Coloured Progressive Matrices.* Bratislava, Slovakia: Psychodiagnosticke a Didakticke Testy.

Floud, R., Wachter, K., & Gregory, A. (1990). *Height, Health, and History.* Cambridge, England: Cambridge University Press.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95,* 29-51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171-191.

Flynn, J. R. (2000). IQ gains, WISC subtests and fluid ***g***: ***g*** theory and the relevance of Spearman's hypothesis to race. In G. R. Bock, J. A. Goode, & K. Webb (Eds.), *The Nature of Intelligence (Novartis Foundation Symposium 233)* pp. 202-227. Chichester, England: Wiley.

Garfinkel, R., & Thorndike, R .L. (1976). Binet item difficulty: Then and now. *Child Development, 47,* 959-965.

Greenfield, P .M. (1998). The cultural Evolution of IQ. In U. Neisser (Ed.), *The Rising Curve* (pp. 81-124). Washington, DC: American Psychological Association.

Heron, A., & Chown, S. (1967). *Age and Function.* London: Churchill.

HMI (Scotland) (1980). *Learning and Teaching in Primary 4 and Primary 7.* Edinburgh: HMSO.

Hoffman, H. V. (1983). *Regression Analysis of Test Bias in the Raven's Progressive Matrices for Anglos and Mexican-Americans.* Unpublished doctoral dissertation, Department of Educational Psychology, Graduate College, University of Arizona.

Hoffman, H. V. (1990). In J. Raven et al., A *Compendium of North American Normative and Validity Studies, Research Supplement No.3 to the Manual for Raven's Progressive Matrices and Vocabulary Tests* (pp. 21-31). San Antonio, TX: Harcourt Assessment.

Hollingshead, A. B. (1967). In C. M. Bonjean, R. J. Hill, & S. D. McLemore (Eds.), *Sociological Measurement*: *An Inventory of Scales and Indices.* San Francisco: Chandler.

Holmes, B. J. (1980). *British Columbia Norms for Wechsler Intelligence Scale for Children-Revised; Peabody Picture Vocabulary Test; Slosson Intelligence Test; Standard Progressive Matrices, and Mill Hill Vocabulary Scale.* Unpublished manuscript, University of British Columbia: Faculty of Education.

Hyman, H. H. (1955). *Survey Design and Analysis*: *Principles, Cases, and Procedures.* Glencoe, IL: Free Press.

Jackson, P. W. (1986). *The Practice of Teaching.* New York: Teachers College Press.

Jaworowska, A., & Szustrowa, T. (1991). *Podrecznik Do Testu Matryc Ravena.* Warsaw: Pracownia Testow Psychologicznych Polskiego Towarzystwa Psychologicznego.

Jensen, A. R. (1974). How biased are culture-loaded tests? *Genetic Psychology Monographs, 90,* 185-244.

Kahn, H., Spears, J., & Rivera, L. (1977). *Applicability of Raven Progressive Matrices Tests with School Children in Puerto Rico.* Hato Rey, Puerto Rico: Department of Education.

Kratzmeier, H., & Horn, R. (1979). *Manual: Raven-Matrizen-Test, Standard Progressive Matrices.* Weinheim, Germany: Beltz Test.

Lovaglia, M. J., Lucas, J. W., Houser, J. A., Thye, S. R., & Markovsky, B. (1998). Status processes and mental ability test scores. *American Journal of Sociology, 1,* 195-228.

Martinolli, L. (1990). *Etude et Reetalonnage des Matrices Progressives Couleur*. Unpublished manuscript, L'Institute de Psychologie, Universite de Fribourg.

Martorell, R. (1998). Nutrition and the worldwide rise in IQ scores. In U. Neisser (Ed.), *The Rising Curve*. Washington, DC: American Psychological Association.

McGillicuddy-DeLisi, A. V. (1985). The relationship between parental beliefs and children's cognitive level. In I. E. Sigel (Ed.), *Parental Belief Systems: The Psychological Consequences for Children* (pp. 261-299). Hillsdale, NJ: Erlbaum.

McGillicuddy-DeLisi, A. V., DeLisi, R., Flaugher, J., & Sigel, I. E. (1987). Family influences on planning. In S. L. Friedman, E. K. Scholnick, & R. R. Cocking (Eds.), *Blueprints for Thinking: The Role of Planning in Cognitive Development* (pp. 395-427). New York: Cambridge University Press.

Mehlhorn, H. G. (1980). Aspekte der Geistigen Entwicklung Jugendlicher. In W. Freidrich und H. Muller (Eds.), *Zur Psychologie der 12 bis 22 Jahrigen.* Berlin, Germany: VEB Deutscher Verlag der Wissenschaften.

Miao, E. S. Y. (1993). Translation of J. Raven, J. C. Raven, & J. H. Court, *Manual for Raven's Progressive Matrices Tests.* Taiwanese Edition. Taiwan: Chinese Behavioural Science Corporation.

Miao, E .S. Y., & Huang, W. (1990, July). *Standardisation and validation of CPM, SPM, and APM in Taiwan, ROC.* Paper presented to 22nd International Congress of Applied Psychology, Kyoto, Japan.

Miller, K. A., Kohn, M. L., & Schooler, C. (1985). Educational self-direction and the cognitive functioning of students. *Social Forces, 63,* 923-944.

Miller, K. A., Kohn, M. L., & Schooler, C. (1986). Educational self-direction and personality. *American Sociological Review, 51,* 372-390.

Neisser, U. (Ed.) (1998). *The Rising Curve.* Washington, DC: American Psychological Association.

New Zealand Council for Educational Research. (1984). *Standard Progressive Matrices: New Zealand Norms Supplement.* Wellington, New Zealand: NZCER.

Nidich, S. I., Morehead, P., Nidich, R. J., Sands, D., & Sharma, H. (1993). The effect of the Maharishi Student Rasayana Food Supplement on non-verbal intelligence. *Personality and Individual Differences, 15*, 599-602.

Nickerson, R., Perkins, D. N., & Smith, E. (1985). *The Teaching of Thinking.* Hillside, NJ: Erlbaum.

Owen, K. (1992). The suitability of Raven's Standard Progressive Matrices for various groups in South Africa. *Personality and Individual Differences, 13*, 149-159.

Owens, W. A. (1966). Age and mental abilities: A second adult follow-up. *Journal of Educational Psychology, 57,* 311-325.

Pollitt, E., & Saco-Pollitt, C. (1996). On the role of the physical environment in the development of intelligence. In D. K. Detterman (Ed.), *Current Topics in Human Intelligence: Volume 5: The Environment.* Norwood, NJ: Ablex.

Plomin, R. (1989). Environment and genes. *American Psychologist, 44*(2), 105-111.

Plomin, R., & Daniels, D. (1987). Why are children in the same family so different from one another? *Behavioral and Brain Sciences, 10,* 1-15.

Raven, J. (1980). *Parents, Teachers and Children: An Evaluation of an Educational Home Visiting Programme.* Edinburgh: Scottish Council for Research in Education.

Raven, J. (1981). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.1, The 1979 British Standardisation of the Standard Progressive Matrices and Mill Hill Vocabulary Scales, Together With Comparative Data From Earlier Studies in the U.K., US, Canada, Germany, and Ireland.* San Antonio, TX: Harcourt Assessment.

Raven, J. (1987). Values, diversity, and cognitive development. *Teachers College Record, 89,* 21-38.

Raven, J. (1989). The Raven Progressive Matrices: A review of national norming studies and ethnic and socioeconomic variation within the United States. *Journal of Educational Measurement, 26,* 1-16.

Raven, J. (1995). Methodological problems with the 1992 standardisation of the SPM: A response. *Personality and Individual Differences, 18*(3), 443-445.

Raven, J. (2000a). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.3 (Second Edition): A Compendium of International and North American Normative and Validity Studies Together with a Review of the Use of the RPM in Neuropsychological Assessment by Court, Drebing, & Hughes*. San Antonio, TX: Harcourt Assessment.

Raven, J. (2000b). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1-48.

Raven, J., Johnstone, J., & Varley, T. (1985). *Opening the Primary Classroom.* Edinburgh: Scottish Council for Research in Education.

Raven, J., Raven, J. C., & Court, J. H. (1998a, updated 2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1: General Overview*. San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (1998b). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 2, The Coloured Progressive Matrices.* San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (1998c). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 4, The Advanced Progressive Matrices.* San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (1998d). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 5: The Mill Hill Vocabulary Scale.* San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3: The Standard Progressive Matrices*. San Antonio, TX: Harcourt Assessment.

Raven, J. C. (1941). Standardisation of Progressive Matrices, 1938. *British Journal of Medical Psychology, XIX,* Part 1, 137-150.

Raven, J. C., Court, J. H., & Raven, J. (1995). *Raven, Matrices Pogresivas (Escalas*: *CPM, SPM, APM).* Madrid: TEA Ediciones, S.A.

Raven, J. C., & Walshaw, J. B. (1944). Vocabulary tests. *British Journal of Medical Psychology, 20,* 185-194.

Sahin, N., & Duzen, E. (1994). *Turkish Standardization of the Raven's SPM (Ages 6-15)*. Paper presented at the 23rd International Congress of Applied Psychology, 17-22 July, Madrid, Spain.

Scarr, S., Webber, P. L., Weinberg, R. A., & Wittig, M. A. (1981). Personality resemblance among adolescents and their parents in biologically related and adoptive families. *Journal of Personal Social Psychology, 40,* 885-898.

Schaie, K. W. (Ed.). (1983). *Longitudinal Studies of Adult Psychological Development.* New York: Guilford Press.

Schaie, K. W. (1994). The course of adult intellectual development. *American Psychologist, 49*(4), 304-313.

Schoenthaler, S. J., Amos, S. P., Doraz, W. E., Kelly, M. A., & Wakefield, J. (1991). Controlled trial of vitamin-mineral supplementation on intelligence and brain function. *Personality and Individual Differences, 12(4)*, 343-350.

Schoenthaler, S. J., Amos, S. P., Eysenck, H. J., Peritz, E., & Yudkin, J. (1991). Controlled trial of vitamin-mineral supplementation: Effects on intelligence and performance. *Personality and Individual Differences, 12(4)*, 351-362.

Schooler, C., Mulatu, M. S., & Mesfin, S. (2001). The continuing effects of substantively complex work on the intellectual functioning of older workers. *Psychology and Aging 16(3) 466-482*.

Scott, E. M., Illsley, R., & Thomson, A. M. (1956). A psychological investigation of primigravidae: Part II: Maternal social class, age, physique and intelligence. *Journal of Obstetrics and Gynaecology of the British Empire, LXIII(3)*, 338-343.

Sigel, I. E., & Kelley, T. D. (1988). A cognitive developmental approach to questioning. In J. Dillon (Ed.), *Classroom Questioning and Discussion: A Multi-disciplinary Study.* Norwood, NJ: Ablex.

Sigman, M., & Whaley, S. E. (1998). The role of nutrition in the development of intelligence. In U. Neisser (Ed.), *The Rising Curve*. Washington, DC: American Psychological Association.

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence, Volume 2* (pp. 47-103). Hillsdale, NJ: Lawrence Erlbaum Associates.

Spicher, S. (1993). *Nouvel Etalonnage du SPM-38*: *Raven dans le ciel Fribourgeois*. Unpublished manuscript, L'Institute de Psychologie, Universite de Fribourg.

Stallings, J., & Kaskowitz, D. (1974). *Follow Through Classroom Observation Evaluation 1972-1973.* Menlo Park, CA: Stanford Research Institute. Report URU-7370.

Teasdale, T. W., & Owen, D .R. (1989). Continuing secular increases in intelligence and a stable prevalence of high intelligence levels. *Intelligence, 13,* 255-262.

Thorndike, R. L. (1975). *Mr. Binet's Test 70 Years Later.* Presidential Address to the American Educational Research Association.

Thorndike, R. L. (1977). Causation of Binet IQ decrements. *Journal of Educational Measurement, 14,* 197-202.

Tuddenham, R. D., Davis, L., Davison, L., & Schindler, R. (1958). *An experimental group version for school children of the Progressive Matrices.* Unpublished manuscript, University of California. See also Abstract: *Journal of Consultant Psychology, 22,* 30.

United States Government, Bureau of the Census (1984). *Statistical Abstract of the United States, 1983.* Washington, DC: U.S. Government Printing Office.

Vernon, P. E. (1969). *Intelligence and Cultural Environment*. London: Methuen.

Vodegel-Matzen, L. B. L. (1994). Performance on Raven's Progressive Matrices. Ph.D. Thesis, University of Amsterdam.

Vitz, P. C. (1990). The use of stories in moral development. *American Psychologist, 45*(6), 709-719.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes.* Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1981). The genesis of higher mental function. In J. V. Wertsch (Ed.), *The Concept of Activity in Society Psychology.* Annank, NH: Sharpe.

Webber, R. J. (1977). *The National Classification of Residential Neighbourhoods. PRAG Technical Paper TP23.* London: Centre for Environmental Studies.

Zhang, H.-C., & Wang, X.-P. (1989). Chinese standardisation of Raven's Standard Progressive Matrices. *Psychological Test Bulletin,* November, *2*(2), 36-39. Hawthorn, Victoria: ACER.

Appendix

Sampling Procedures, Sample Sizes and Data Management

This Appendix outlines some of the considerations which have guided our choice of sampling methodology and data analysis and presentation.

Virtually all statistical tests assume that the groups between which it is desired to discriminate, or from which it is proposed to generalise, are random samples from some wider population. Yet, while attaching much importance to sophisticated statistical technique, psychologists rarely examine the quality of their samples. It is not uncommon for them to assume, for example, that results obtained in studies of psychology students will apply to all people or all people in a category – such as males and females.

Commonly, even when an effort is made to ensure that a population tested is representative of some wider population, "quota sampling" techniques are employed. In these an effort is made to ensure that the demographic characteristics of the population tested correspond to those of some wider population to whom it is hoped to generalise.

Yet, even by the time Hyman wrote his classical book on *Survey Design and Analysis* (1955), it had been repeatedly demonstrated that not only do opinion polls based on huge numbers yield much less accurate data than studies based on much smaller, but randomly selected, samples, so, too, do studies based on quota samples.

For these reasons we have, in our own work, sought to employ systematic random sampling procedures, wherever possible doing so within strata which have been chosen to yield the correct proportions in certain demographic categories required to correspond to wider demographic statistics.

It is important to note that stratification via demographic statistics is a very different matter to asking individual researchers to locate and test specified numbers of people within a number of categories identified in terms of such things as sex, age, socioeconomic status, and ethnic group.

In the 1979 British study we were able, with the aid of funds from the Social Science Research Council, to conduct the study in seven areas of the country which previous research (Webber, 1977) had shown to cover the main variance within the country while at the same time being collectively representative of the country as a whole. We were even able

to over-sample particular areas in order to have large enough numbers of respondents to permit detailed comparisons between areas and then re-weight the data to produce have the correct effect when combined with other data in the overall statistics.

In most of the other work summarised in this chapter this has not been possible. It has been necessary to work with collaborators who were interested in contributing to the study and to do what was feasible under the circumstances. As far as possible, we have sought both (1) areas with demographically balanced populations, and (2) a *range* of areas located in parts of the country having very different demographic characteristics. Within areas we have tried to ensure that the samples tested were selected using some strictly random method. In some cases complete lists of names have been obtained and then sampled using a random start and a fixed sampling interval. In other cases it has been necessary to compromise by doing such things as systematically select buildings and classrooms within school districts to be representative of the whole and then test all the children in those classrooms. Such clustering pushes up the numbers but it does not, in fact, yield better samples.

Naturally, data obtained in these ways – unlike those obtained from the 1979 U.K. standardisation among young people – cannot be pooled using routine statistical procedures. Instead, it has to be combined making due allowance for deficiencies in the data set and giving more weight to the more balanced and complete samples.

Chapter 9

Does the "Flynn Effect" Invalidate the Interpretation Placed on Most of the Data Previously Believed to Show a Decline in Intellectual Abilities with Age?

Francis Van Dam and John Raven

Acknowledgements

The authors are deeply indebted to Jean Raven for her continuous assistance. Also to the academic authorities of Louvain University, at the time of the first test, Mrs R. Berte (Faculty of Psychology and Educational Sciences), Mrs M. Tits, Professors L. Ghosez and I. De Aguirre (Faculty of Sciences), but also for their aid to the feasibility of the retest, Professor M. Crochet, Rector, and Professor M. Hupet, Decan of the Faculty of Psychology and Educational Sciences. Internal services of the University and Mr P.Escoyez deserve our sincere gratitude.

Assistance of Mrs. L. Sohy and Messrs. Tran Quoc Duy and J.-C. Van Dam for statistical controls and graphical versions of our data has been highly appreciated.

Abstract

This paper is divided into two Parts. *Part I* presents longitudinal data on the way people's performance on the *Advanced Progressive Matrices* changed as they aged. The data come from 99 Louvain University students who were first tested in 1970 or 1971 and followed up in 2002-03. The data relate directly to the questions about the validity of widely held beliefs about the "decline" in ability with increasing age that have been called into question by data published by Flynn and others who have documented a dramatic increase in scores with date of birth. These

data documenting the secular increase in RPM scores with date of birth suggest that the age differences in earlier cross-sectional studies may need to be reinterpreted. Instead of showing that scores decline with age they may simply show that people born more recently get higher scores. The data reported in this paper suggest that most of the apparent "decline" is probably due to the secular increase in scores. However, they also show that trends with age are far from universal. Some respondents' scores increased almost as much as others declined. *Part II* presents data which bear on the question of whether, as people age, it is not so much their ability to solve problems as the time they need to do so that declines.

Part I: Changes In The Number of Problems Correctly Solved As People Grow Older

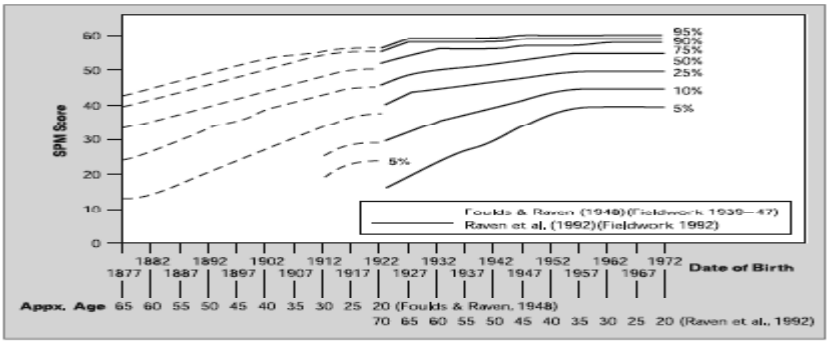
1. Introduction

Data on the "intelligence" test scores achieved by people of various ages that have been assembled by testing samples of the general populations of a number of countries at a particular time (e.g. Deary, 2000) – i.e. data from *cross-sectional* studies – have generally been interpreted to mean that most abilities decline steadily as age advances. It has usually been claimed that "fluid intelligence" or "reasoning ability" declines most rapidly and routine skills and knowledge decline less. A striking graph supporting this position will be found in Schaie and Willis (1986).

Flynn (1984, 1987), on the other hand, assembled data collected mainly from 18 year olds entering military service in a wide range of countries and demonstrated that there had been a huge inter-generational increase in scores. Similar data had earlier been published by such authors as Thorndike (1975, 1977), Garfinkel and Thorndike (1976), and Bouvier (1969) and Raven (1981). While discussing his data, Flynn suggested that at least a significant proportion of what had previously been believed to be a "decline with age" in reality reflected an increase with date of birth (because the older people in the cross-sectional samples had, of course, been born earlier).

This speculation was strongly reinforced by Raven (1998, 2000). In 1992 he conducted a cross-sectional norming study with the *Standard Progressive Matrices* on a sample which might be taken to yield results that would approximate those that would be obtained from testing a

Figure 9.1. *Standard Progressive Matrices*
100 Years of Eductive Ability



cross-section of adults in the UK and compared the graphed percentile age norms from that study with similar norms estimated by his father from data available 50 years earlier – i.e. about 1942 (Raven, J.C., 1948; Foulds, & Raven, J. C., 1948). In Figure 9.1 (reproduced from Raven, 2000) the cross sectional norms from the two studies, which are normally plotted with increasing age as the X axis (thus purporting to show a "decline" with age), are plotted by date of birth. The results from each study thus show an increase in scores with date of birth until the dates on which the two cross sectional studies were conducted. If the apparent increases with date of birth were, in reality, due to declines with age then the two sets of graphs (from the 1942 and 1992 studies) would be side by side, each set starting from approximately the same point. But that is not what we see. Instead, the graphs show a continuous increase. Particularly striking is the fact that the scores of a sample of the 1922 birth cohort tested circa 1942 (when they were approximately 20 years of age) were almost the same as those of a sample of the same birth cohort tested in 1992 (when they would have been approximately 70 years of age).

It may be noted in passing that the data are similar to Flynn's in that they present results obtained at different points in time, but differ in the sense that, whereas Flynn's data related mainly to 18 year old military conscripts, these data cover all age groups or, stated differently, and perhaps in a way more relevant to the theme of this article, a cross-section of people from every birth cohort from 1877 to 1972.

In any event, it would seem from these data that *Raven Standard Progressive Matrices* scores at all levels of ability do *not* decline with age but, instead, increase steadily with date of birth.

The significance of these data cannot be over-estimated. If the absence of almost any decline with age were confirmed, the implication would be that the interpretation placed on the data collected in the course of the numerous cross-sectional studies already mentioned would have to change.

To fully test this hypothesis, it would be necessary to draw together the results of several longitudinal studies in which cross-sections of the adult population had been tested on a number of occasions.

Because of the difficulties involved in tracing people who had been tested 20 or more years earlier, very few such studies exist – with *any* test, let alone the Raven Progressive Matrices.

One of the most widely cited studies – and one which is generally believed to have contributed such data – is Schaie's (1983) *Seattle*

Longitudinal Study, which was conducted with Thurstone's *Primary Mental Abilities* test. Yet even this study was plagued by huge population losses and relatively short follow-up periods. (Although the study started out with 500 respondents, this number had fallen to 92 after 28 years – i.e. after much the same time lapse as the period over which the present study was conducted.) These problems forced Schaie to make endless adjustments to his basic data in order to arrive at what might, at best, be viewed as extremely tentative conclusions.

Owens (1959, 1966) examined the scores of 129 college freshmen who had been tested with the *Army Alpha Examination* at ages 19, 50 and 61. From 19 to 50 years of age there was little change in mean scores, and, even between 50 and 61 years of age, the only significant decline was on numerical ability and this amounted to a mere .20 of a standard deviation.

Schwartzman et al. (1987) also reported a decline of only .85 of a standard deviation in non-verbal ability over a 40-year interval time period among 260 respondents whose average age at retest was 64.7.

Pushkar Gold and collaborators (1991), quoted by Deary (2000), also indicate a decline in so-called "fluid" (but not "crystallised") "intelligence" in their longitudinal data based on 300 respondents between 25 and 65 years.

Deary and his colleagues have published a series of remarkable studies (of which Deary *et al,* 2000 is but one example) based on a follow up, in 1998, of a sample of 101 Scottish adults from the 1921 birth cohort, the entire population of whom had been tested in 1932 when they were 10-11 years old. Unfortunately, because the sample had not been tested between these two time points, there were no data on the scores obtained at maturity (i.e. at about 18 years of age) so it was impossible to find out whether there had been any development or decline in scores with advancing age. Nevertheless, it is worth reporting that the correlation between the scores of these people tested 66 years apart was 0.63 (corrected for restriction of range 0.73).

In order to check the validity of the now strengthening hypothesis that the data that had previously been thought to show a decline in scores with age really revealed an increase in scores with date of birth, the second author (Raven) contacted a number of researchers who were known to have tested large samples on one or other of the *Raven Progressive Matrices* tests a number of years ago and who might have kept sufficient information on the names and addresses of those concerned to make it feasible to consider recontacting them.

The primary author of this article was one of those so contacted. In the course of a research he conducted in 1970-71 he had tested 1,095 first year students in the University of Louvain on the *Advanced Progressive Matrices* and it appeared that, through the alumni records of the university, it might be possible to trace the current addresses of some of those concerned. The difficulties involved in doing this and the overall success or otherwise of the operation will be described below.

2. The Design of the Present Study

2.1. Participants

1095 first year students in the French Speaking division of the University of Louvain had been tested with a 48-item version of Set II of the *Advanced Progressive Matrices* (then known as *PMA 1947*) in 1970 and 1971.

For the present study, the scores of all Asiatic, African, Australian, and similar students were discarded, leaving only Belgian, French, German, and similar respondents: 844 in all.

In 1970 and 1971, students were tested, in groups of about 50, soon after their registration. They came mainly from the Faculties of Applied Sciences. There were no students of the exact sciences or Arts. No longitudinal study was planned at this time. The research was intended to find ways of minimising academic failure and assist in career choice. The testing was organised with the support of the University authorities. In 1970 the testing was voluntary but it was compulsory in 1971.

2.2. Tests and testing procedures in 1970 and 1971

The test used was the original version of what has, from 1962, been known as the *Advanced Progressive Matrices* (APM). Then known as *PMA 47,* this had an introductory, practice, set of 12 items (termed "Set I") followed by a main test ("Set II") of 48 items. However, the introductory set (Set I) was not administered in 1971. Set II was administered with a 40 minute time limit. In both years, respondents were asked to indicate the item in Set II that they were working on after 30 minutes. Unfortunately, not all students did so.

2.3. Overall results

The mean score of the 1970 group (n = 288) was 34.25 and that of the 1971 group (n = 556) 35.48. The distribution of scores for the 1971

group (to whom Set I was not administered) was slightly less Gaussian than that of the 1970 group.

The scores of both groups were well above ($t = 5.2$, $p = .001$) those for secondary school students from the same region at the same period ($M = 32.21$, $n = 278$) (Florquin, 1964).

2.3.2. Variation in scores with gender and subject studied

Overall, there were no differences between the mean scores of men (34.83) and women (35.36). However this was not true within subjects. Women studying in faculties in which they were a distinct minority (such as veterinary medicine and agronomics) got higher scores than men studying those subjects. They had lower scores in other subjects, such as biology, where they were in a majority. The impossibility of generalising from a population with such internal variation to sex differences in general is thus immediately obvious and it invalidates any attempt to tease out any generalisable differential effects of gender from the changes in scores over time to be documented below.

3. The Follow-Up Study

3.1. Participants.

The ex-students were traced through the University alumni records. It was possible to trace only 217 of the 844 students initially tested 30 years earlier, and this with the utmost difficulty. These were sent a letter, accompanied by a test booklet, instructions, an answer sheet, and a prepaid return envelope. Only 99 (60 males and 39 females) returned the material. The proportion of men and women happened to be the same as in the initial study (respectively. 61% and 39 %). 57 of the respondents from whom replies were obtained had first been tested in 1970 and 42 in 1971.

3.2. Procedure.

As mentioned, those ex-students whose addresses could be traced were sent a letter explaining the purpose of the study (and reminding them of the earlier testing) and asking them to again try to solve the problems of the same (48-item) version of APM Set II that they had taken in 1970 or 1971. At this time they were not given any information about the scores they had obtained on the previous testing, but they were told they would obtain feedback on the scores obtained on both occasions once the new scores were obtained.

Table 9.1. *Advanced Progressive Matrices, Set II, 48 Item Version*
Correlations Between 1970/71 and 2002-03 Scores;
Correlations Between Number of Items Answered Correctly (SCORE) and Number of Items Attempted (SCAN)

	Initial Test (1970-1971)				Retest (2002-2003)			
	Score 30 min	**Scan 30 min**	**Score 40 min**	**Scan 40 min**	**Score 30 min**	**Scan 30 min**	**Score 40 min**	**Scan 40 min**
Test								
Score 30 min	1.00							
Scan 30 min	0.59	1.00						
Score 40 min	0.89	0.38	1.00					
Scan 40 min	0.39	0.78	0.29	1.00				
Retest								
Score 30 min	0.60	0.38	0.49	0.23	1.00			
Scan 30 min	0.39	0.36	0.23	0.19	0.77	1.00		
Score 40 min	0.57	0.29	0.50	0.15	0.92	0.58	1.00	
Scan 40 min	0.22	0.29	0.18	0.08	0.59	0.85	0.52	1.00

Based on the 78 respondents who indicated the number of the problem they were working on at the end of both 30 and 40 minutes on both occasions.

Ex-students who had not completed their University studies – i.e. who had failed to graduate – were sent the same package of materials as those who graduated.

In the instructions for taking the test, respondents were asked to limit the time they worked on it to 40 minutes. Also, to circle on the answer sheet the number of the problem they were working on after 30 minutes had elapsed. Only 78 of the 99 respondents who completed the retest marked the item they had reached after 30 minutes at both the initial and follow up testing. As a result, some of the results reported below refer to the overall sample (n= 99) while others are derived from the more restricted sample of 78.

Once the test booklet had been returned and the answer sheet scored, respondents were sent simple information about their performance at both the first and second occasion. The letter was written in a rather neutral way, reporting score(s) in terms like "average, superior, highly superior, inferior to the mean" etc. Pessimistic judgments were avoided as much as possible, but an extended commentary explored possible implications for work and career. A lottery ticket was enclosed with the report, which was also a token of thanks for participation in the study.

4. Results

4.1. Test-retest reliability.

In addition to analysing the number of items answered correctly after 30 and 40 minutes, an analysis was also made of the total number of items *attempted* – i.e. how far into the test respondents had got by the time they reached these two time markers. The latter scores are referred to as SCAN scores (number of attempted or *scanned* items). Although discussed later, Table 9.1 includes correlational statistics for these scores as well as for the "number correct" scores (SCORE), obtained at the first and second testing.

As can be seen from Table 9.1, the overall test-retest correlation between the scores achieved at the end of 40 minutes in the initial and follow up testing was 0.50. In other words, the initial scores explained only 25% of the variance in scores 30 years later. This is less than was the case in Deary's study (which was conducted with a different test over a longer time period). Part of the difference undoubtedly stems from the restricted range of abilities covered by the present study, and allowance must also be made for this when interpreting the results presented below.

Although we have not included them here, we calculated these correlations broken down by sex and year of first testing. From this more detailed analysis it emerged that the test-retest correlations for the final scores obtained in 40-minutes were 0.46 for females and 0.55 for males. Also that the overall correlation was lower (0.36) for the group first tested in 1970 and higher (0.62) for the group first tested in 1971.

From Table 9.1 it will be seen that, among the 78 respondents who recorded the number of the item they were working on at the end of 30 minutes in both 1970/71 and 2002/03, the overall test-retest correlation between the scores achieved after 30 minutes work was 0.60 and that this was higher than the test-retest correlation on completion of the test at the end of 40 minutes (which was 0.50).

The correlation between the number of items attempted at first and second testing was .36 at the end of 30 minutes and .08 after 40 minutes had elapsed. It follows that there seem to have been some fairly dramatic changes in what might be considered to be an index of people's speed of work over the intervening 30 years and that this was greater than the changes in the number of items answered correctly.

Figure 9.2. *Advanced Progressive Matrices, Set II, 48 Item Version, 40 Minute Time Limit*
Distribution of 1970/71 Scores for 99 Respondents Retested in 2002-03

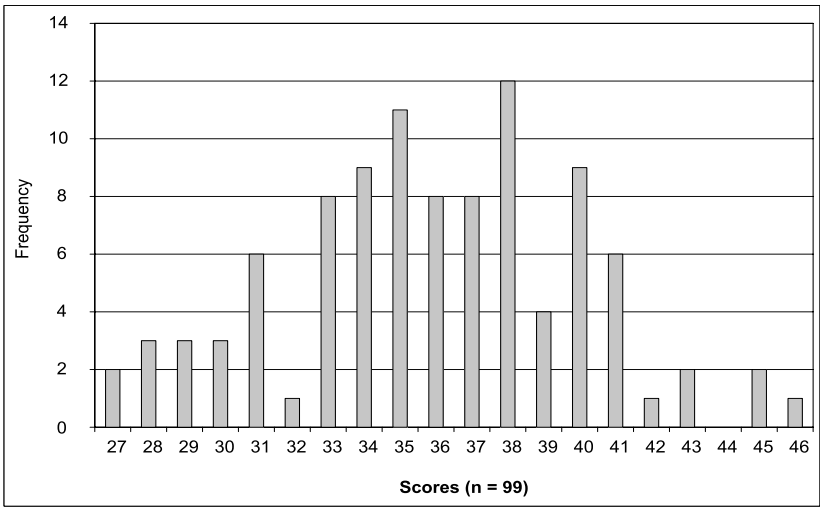


Figure 9.3. *Advanced Progressive Matrices, Set II, 48 Item Version, 40 Minute Time Limit*
Distribution of Final (40 Minutes) Scores Obtained in 2002/03 Compared with 1970-71 Distribution (n = 99)

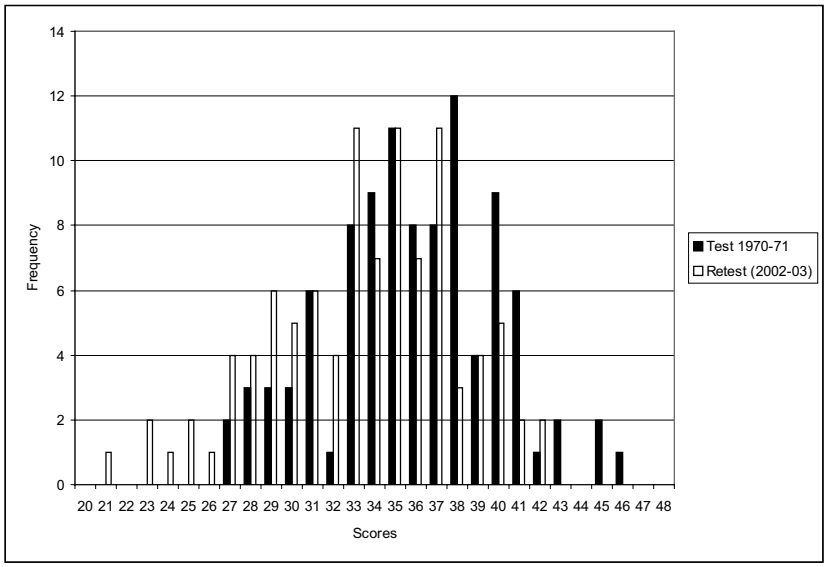
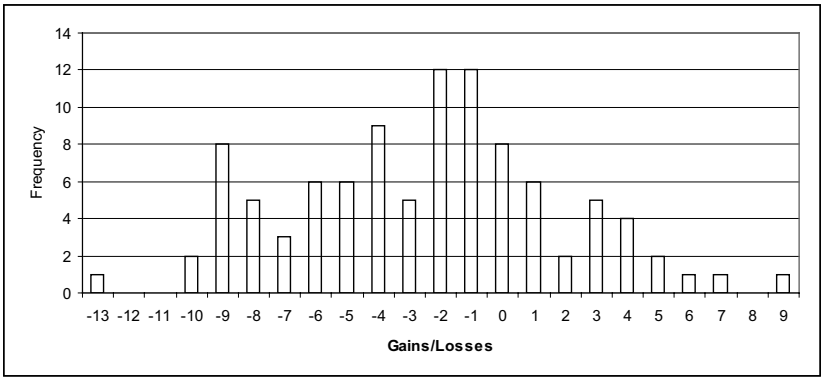


Figure 9.4. ***Distribution of Losses and Gains 1970/71-2002/03: 40 Minute Time Limit (n = 99)***
A Negative Number Means That The Scores Have Gone Down From 1970/71 to 2003/04



4.2. Changes over time

4.2.1. Overall scores

The average final score in the overall time limit of 40 minutes for the whole group (99 respondents) fell from 35.9 to 33.4 between 1970/71 and 2002/03. The standard deviation increased slightly, from 4.13 to 4.57. (The overall distribution of the 844 students tested in 1970/71 was Gaussian, according to the chi-square formula.) A histogram of the distribution of the final (40 minute time limit) 2002/03 scores for the 99 respondents involved in the follow up study is given in Figure 9.1, and, in Figure 9.2, the same histogram is compared with the 1970/71 histogram of the same 99 S's. A histogram of the individual gains and losses is presented in Figure 9.3.

The t test for paired samples between individual scores at both tests equals 5.726 (df = 98) and yields a very highly significant difference ($p < 0.000$).

In addition to noting the overall decrease in scores, it is important to draw attention to the great individual variation in the change scores: almost as many scores have increased as decreased, none have fallen by more than 13 points and the majority changed very little.

Attention should be drawn to the fact that it is more difficult to increase scores than to decrease them because increasing them involves solving more difficult items. Nevertheless, from the point of view of comparing relative gains and losses, this is actually a minor problem. Much more basic problems stem from the fact that there is no guarantee that the items in the test are equally spaced in terms of difficulty. Consequently, as Prieler and Raven have shown in another chapter of this book, assessing relative gains and losses is fraught with difficulty.

4.2.2. Differential changes among those first tested in 1970 compared with those first tested in 1971

The mean scores of those first tested in 1970 fell less than those of the group first tested in 1971. The scores of the first group (n = 41) declined by only 1.27 (from 34.39 to 33.12) points although the standard deviation of scores for this group increased considerably (from 3.8 to 4.9). The average decline in the scores of those first tested in 1971 (n = 58) was greater (3.38 points, from 36.97 to 33.59) but there was a smaller increase in the S.D. (from 4.18 to 4.43). However, since the original mean score of the first group was already relatively low (34.39), it means

that those respondents had never been able to solve the more difficult items that the second group solved on the first occasion but became unable to reach or solve within the time limit on the second occasion. The change scores may therefore have different meanings for the two groups.

4.2.3. Gender differences.

Overall, the full-time scores of the 60 male respondents declined less than those of the 39 females, the declines in the means being respectively 1.78 and 3.62 (Figure 4). Both declines are statistically significant ($t = 3.1$ $df = 59$ $p < .003$ for men and $t = 5.6$ $df = 38$ $p < 0.000$ for women). This gender difference is all the more striking in that there was no significant difference between the scores of 39 female and 60 male participants when tested in 1970/71 (M = 36.26 for men and 35.67 for women).This is consistent with the absence of significant difference stated between the whole samples of men ($n = 517$) and women ($n = 327$) at their first testing in 1970-1971.

Despite the impression given by these overall figures, the results appear to vary considerably with whether one is dealing with the group initially tested in 1970 or 71 (whom, it will be recalled, were subject to

Figure 9.5. *Advanced Progressive Matrices, Set II, 48 Item Version, 40 Minute Time Limit*
Mean Scores of Men (n = 60) and Women (n = 39) at Test (1970-71) and Retest (2002-03)

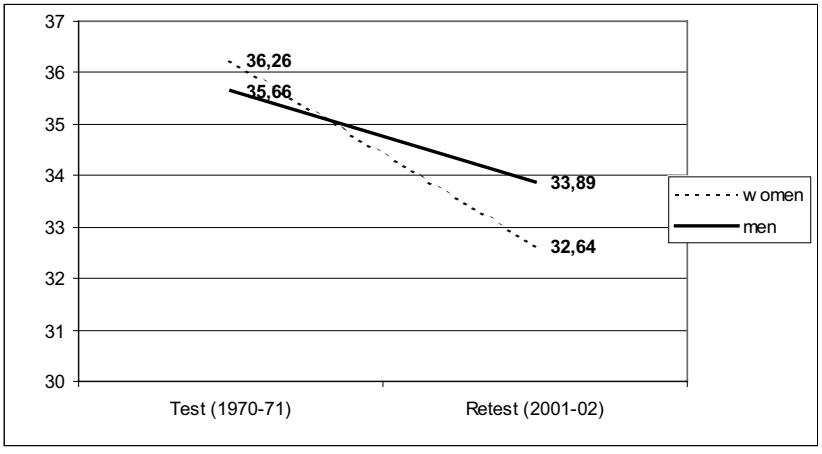


Table 9.2. *Advanced Progressive Matrices, Set II, 48 Item Version, 40 Minute Time Limit* **Initial and Retest Scores at 30 and 40 Minutes by Date of Initial Testing and Gender**

30 Minutes	**Gender**	**Test (1971-72)**	**Retest (2002-03)**	
1970		*2002*		*Difference*
	Female	M = 33.18	M = 28.09	-5.09
	(*n* = 11)	s.d = 3.40	s.d = 4.32	+0.92
	Male	M = 32.96	M = 29.96	-3.00
	(*n* = 25)	s.d = 3.30	s.d = 4.99	+1.69
	Total	M = 33.03	M = 29.39	-3.64
	(*n* = 36)	s.d = 3.28	s.d = 4.81	+1.53
1971		*2002*		*Difference*
	Female	M = 33.88	M = 30.47	-3.41
	(*n* = 17)	s.d = 5.18	s.d = 3.89	-1.29
	Male	M = 34.68	M = 30.08	-4.60
	(*n* = 25)	s.d = 4.92	s.d = 5.89	+0.97
	Total	M = 34.36	M = 30.24	-4.12
	(*n* = 42)	s.d = 4.98	s.d = 5.13	+0.15
40 Minutes				
1970		*2002*		*Difference*
	Female	M = 35.18	M = 31.91	-3.27
	(*n* = 11)	s.d = 3.74	s.d = 4.64	+0.90
	Male	M = 34.24	M = 33.00	-1.24
	(*n* = 25)	s.d = 3.68	s.d = 4.95	+1.27
	Total	M = 34.53	M = 32.67	-1.86
	(*n* = 36)	s.d = 3.67	s.d = 4.82	+1.15
1971		*2002*		*Difference*
	Female	M = 37.29	M = 33.24	-4.05
	(*n* = 17)	s.d = 3.60	s.d = 3.33	-0.27
	Male	M = 37.20	M = 33.72	-3.48
	(*n* = 25)	s.d = 4.69	s.d = 5.05	+0.36
	Total	M = 37.24	M = 33.52	-3.72
	(*n* = 42)	s.d = 4.24	s.d = 4.39	+0.15

different testing procedures). A breakdown of the results by year of first testing and gender is therefore given in Table 9.2 for the 78 respondents who actually indicated which item they were working on at the end of both 30 and 40 minutes in both 1970/71 and 2002/2003.

By and large, the mean scores of women decline more than those of men, irrespective of the initial year of testing. This could be an effect of not practicing a job, as P.E.Vernon (1947) suggested. This hypothesis finds some support in the fact that the largest decline in our sample came from a woman who had never practised any professional job but raised an exceptionally big family.

° ° ° ° °

This study was conducted to throw some light on the question of whether there is still any reason to believe that reasoning – or meaning-making – ability declines with advancing age when the evidence from the cross-sectional data that have in the past been used to justify this claim has, at least to a great extent, been discredited and shown (like similar data on height and life expectancy) to reveal primarily a secular increase in scores with date of birth.

To investigate this hypothesis, it is necessary to compare the magnitude of the, at best far from universal, "decline" in scores with age documented above with the increase over time that would have been expected to have occurred among respondents of similar ability over the same period of time. The comparison has to be with people of similar ability because, as can be seen from Figure 9.1, the original cross-sectional data reveal differential raw score changes over the years at different levels of ability. These have been widely interpreted – for example, by J. C. Raven (1948) himself – as showing that the scores of the more able "decline" "less" than those of the less able. Unfortunately, as can be seen from another chapter in this book, Prieler and Raven (2002) have demonstrated that, obvious as it may seem, such a conclusion cannot really be drawn from these data. Despite this, it is clear from Figure 9.1 that any attempt to find out how changes over time documented through a longitudinal study of a group of people who do not constitute a representative sample of the general population compare with the changes that one would expect to find in data collected in a cross-sectional study conducted at the end of the period must compare like with like. The "decline" to be expected at the 95th percentile would, for example, be very much less than that to be expected at the 5th percentile. In short, we need to compare the (minimal) rates of decline in scores with advancing age documented above with the

magnitude of the changes revealed in the previously mentioned cross-sectional studies for people of similar levels of ability.

Unfortunately, even making such a comparison presents serious difficulties because no good cross-sectional general population norms were available for the *Advanced Progressive Matrices* (*PMA 1947)* in 1970/72. This is partly because the APM was developed with the specific objective of spreading the scores of the top 20% of the general population and would thus not be expected to discriminate within the rest of the population[1].

Nevertheless an undated *Guide to the Use of Progressive Matrices (1947)* published in the UK in about 1948 or 1950 and reprinted one or more times before 1958 (Raven, J. C., c 1950) does include the norms shown in Table 9.3 for 20 year olds for the 48-item version of the test administered with a 40 minute time limit.

The 1963 French Manual (Schutzenberger, 1963) also included two sets of norms.

One set, dated 1960-1962, came from a highly qualified group of 200 respondents aged 25 to 40 comprising engineers, managers and office workers, and reported by the "Services Psychotechniques de la radiotechnique de Suresnes (France)". The group tested appears to be similar in demographic composition to that involved in the present study in that it consisted mainly of graduates who had studied different subjects. The mean APM score was 34.08 with a S.D. of 6.4.

Another standardisation, carried out in 1955, was based on 340 applicants (of both sexes) for jobs with KLM (Royal Dutch Airlines). They appear to be of above average ability in that they were all able to solve more than half of the items in APM Set I.

The 20-24 age group (n = 100) obtained a mean score of 28.1 with a S.D. of 5.3 whereas the 40+ age group had a mean of 22.6 and a S.D. of 6, a lowering which started already for the 35-39 years age group . This decline of 5.5 points in raw scores over 15 years is fairly comparable

Table 9.3. *APM Set II, 48 Item Version, 40 Minute Time Limit*
British Norms for 20 Year Olds
Date of collection/estimation unknown, but probably about 1948-50

Percentile	Score
95	34
90	31
75	26
50	**21**

Table 9.4. *APM Set II, 48 Item Version, 40 Minute Time Limit*
British Cross-Sectional Percentile Norms by Age and Year of Testing

	Age in Years										
	20			30		40		50			
Year	1952	1970 /71	1992	1952	1992	1952	1992	1952	1970 /71	1992	2002 /03
Percentile		Estim.						Estim.	Estim.		Estim.
95	34	39	43	33	43	31	42	30	35	41	43
90	31	36	41	30	41	28	40	26	33	39	43
75	26	31	37	24	37	21	36	19	27	34	37
50	**21**	**26**	**32**	**19**	**32**	**--**	**30**	**--**	**21**	**28**	**31**

Notes: The norms which have, for convenience, been dated circa 1952 come from Raven, J. C. (1952). The norms for the 40 minute version for 20 year olds come directly from that publication, but the norms for 30 and 40 year olds have been estimated from norms for untimed administration. The norms for 50 year olds at that time have been derived by extrapolating the trend in the 20-30-40 year olds' data. The 1970/71 norms were interpolated between the c1952 and 1992 norms.

to that reported at this level of ability in cross-sectional studies conducted with the SPM in the 1950's.

It is clear from the data summarised above that the 1970/71 scores of all 99 respondents involved in the present study were above the 75th percentile of the norms for 20 year olds in the 1950 British norms. They also equal or surpass the 50th percentile from the KLM study and 82 of them (83 %) scored above the 75th percentile from that study. As we have seen, the mean scores matched those obtained in the Suresnes study.

In order to compare the (minor, irregular, and frequently reversed) "decline" in scores with increasing age documented in the present study with that which would have been expected for this ability group on the basis of longitudinal data previously (and erroneously) thought to demonstrate such a "decline", it is necessary to convert the 1992 British norms established for the shorter version of the test published in Great Britain in 1962 (and used from then until the present day) to the equivalents that would have been expected for the 48-item version of APM Set II used in the present study.

Graph 4 in the editions of the (British) *Guide to the Use of the Advanced Progressive Matrices* published between 1965 and 1975 (Raven, J. C., 1965) identifies the items eliminated when Set II was reduced from 48 to 36 items in 1962. Items 1-8 and 17 were eliminated because no one got them wrong. Item 11 was a bad item. And items 44 and 46 were too difficult.

It follows that one can add 8 to scores on the new test that lie between 1 and 2, 9 to scores between 3 and 7, 10 to scores up to 33, and 12 to scores above that in order to obtain the equivalent scores that would have been obtained on the 48 item test. This conversion has been used to prepare Table 9.4.

It seems clear from the data in Table 9.4 and the French data presented above that the mean score of the 20 year old students tested in Louvain in 1970/71 (35.9) was somewhere around the 90th percentile when judged against appropriate norms.

30 years later, that is to say by the time of the follow up study, these ex-students were approximately 50 years old. By then, their average score had fallen to 33.4. This approximates the 90th percentile score of their parents, who might be assumed to have been about 50 years old in 1970/71.

However, the data in the table are more than a little puzzling. The best estimate for the 90th percentile for 50 year olds in a cross sectional

sample conducted in 2002/03 was 43, which is way above the mean score of the ex-Louvain students. Yet these *are* the birth cohort who had been 20 in 1970/71. How could the 90th percentile score of that same birth cohort possibly have increased from 36 to 43 as they got 30 years older? It is all very well to argue that, over the years, there has been an intergenerational increase of this magnitude from generation to generation. It is quite another to believe that, as people age they get much higher scores. In fact the scores of these ex-students did *not* go up. They went *down* and by precisely the amount that would have been predicted from the estimated cross sectional norms for 1970/71. How can it possibly be that the cross-sectional norms for the same birth cohort show such an increase over the years?

The 75th percentile for 20 years olds in the UK data for the 36 items test increased from 14 to 27 from 1962 to 1992, a gain of 13 points, so the same raise can be expected to be 16 on the 48 item test.

In the 1992, 36-item, version of the test and in the context of 20 to 40 years of age, the loss related to maturity equals 1 point (from 27 to 26) for both the UK and the US untimed administrations of Set II (APM Manual 1998 p. 85/86 and 89-91): so there is place for an increase of scores - the Flynn effect - and for a close convergence of the scores of young and mature respondents in the context of recent norms.

If the loss in the 75th percentile norm from a 20 year old in 1962 to a 40 year old in 1992 was only 1 point less (perhaps 2 on a 48 item test), it would suggest that the true decline was about 2 points, which is virtually identical to the actual decline at that level of ability in the Louvain data.

With such figures, it is clear that the bulk of the general population increase has in fact been due to the "Flynn effect".

Part II: Changes in Processing Speed or Strategy

Introduction

Two widely asserted claims in the literature on "intelligence" are (a) that the well-known variance in "intelligence" is mainly attributable to differences in speed of work, reaction time, even rate of neural transmission, and (b) that, as people age, their reaction times slow, thus leading to "declines" in their "intelligence". Deary (2000) has, however, shown that, while no consistent relationships can be established between *Raven Progressive Matrices* scores and most measures of "reaction time", there is a reasonably consistent relationship with "inspection time". Inspection time tests assess how long people need to be 80% accurate in their judgments about which of two lines of markedly different length is the longer. The test itself is not speeded; it is not administered under the stress of time constraints. Prieler and Raven, in another chapter of this book, have shown that timing Item-Response-Theory-based measures of ability seriously contaminates the results because, under timed conditions, many people never reach the more difficult items and therefore cannot demonstrate how well they can do. The result is that the scores obtained constitute an uninterpretable mixture of speed and ability. Forbes (1962) showed that timing the APM (a "power" test constructed to satisfy the requirements of Item Response Theory) seriously discriminates against those who work more slowly and carefully. And both Schaie (see Deary, 2000) and Raven (2000) concluded, together with Deary himself (2000), that the main changes with advancing age stem, not from declines in ability, but in the amount that can be accomplished in a given time.

Given that, in the present study, it was known how many items 78 respondents had attempted both at the end of 30 minutes and on completion of testing at the end of the test, it seemed worthwhile, despite the problems of interpretation introduced by the 40 minute time limit, to review the available data in more detail to see what light could be thrown on the question of whether the primary effect of aging is to slow processing speed rather than to depress ability.

4.3 Retest reliability of the indices

The number of items attempted within a given time might be viewed as an index of processing speed or, at least, some kind of indication of respondents' attitude to risk-taking or trade off between speed and accuracy of work. Unfortunately, the indices of processing speed available from this study (number of items attempted at 30 minutes; number

attempted by the end of the test [40 minutes]; and the difference between the two) are even less reliable (in the statistical sense) than the "number of items answered correctly" scores that have already been examined. For those first tested in 1970, the correlation between the number of items attempted by the end of the test (40 minutes) in 1970 and 2002/03 was 0.07. For those first tested in 1971 this correlation was 0.36. At the end of 30 minutes the respective correlations were 0.02 and 0.57. Clearly, individual differences in the number of items attempted in a given amount of time are unstable. Perhaps this means that they are not a reliable index of "speed of work" or that "speed of work" is itself unstable, i.e. that it changes quite dramatically with ageing or the conditions of test administration. But the change in the correlations over the last ten minutes point at least to an interaction between "speed" and "difficulty". (It may be relevant to point out that, for the *Standard Progressive Matrices*, a number of researchers have found no relationship between final score and the number of items attempted in a given time.)

4.3.1 Number of items attempted in 2002/03 compared with number attempted in 1970/71.

Let's consider the restricted (n = 78) sample of those having marked the item they reached after both 30 and 40 minutes. The number of items attempted by the end of the test (40 minutes) fell on average by more than 4 (44.82 to 40.42) from 1970/71 to 2002/03. At the same time, the variability between respondents – i.e. the standard deviation – increased from 3.30 to 4.97. On average, respondents stopped the retest after attempting 40-41 items, and this irrespective of their initial ability level or of the (former) way of being administered the test, whereas the last attempted item had been, on average, 43.72 (in 1970) and 45.76 (in 1971), giving an overall average for 1970/71 of 45 items examined. In other words, the average number of items attempted after 40 minutes at the retest was just above the average number that had been reached after 30 minutes in the initial test (40.42 vs. 39.58).

4.3.2. Correlations between number of items attempted and number correct in 1970/71 and 2002/03.

The overall correlation between the number of items attempted and number correct was 0.29 in 40 minutes in 1970/71 and 0.52 in 2002/03, but the difference between these correlations is not statistically significant (chi^2 = 3.49 df = 98; $p < .10$).

Table 9.5. *Advanced Progressive Matrices, Set II, 48 Item Version, 40 Minute Time Limit*
Variations in the relationship between number of items attempted (SCAN) and number correct (SCORE) at 30 and 40 minutes (n = 78)

		Test (1970-1971)				Retest (2002 - 2003)			
		30 minutes		40 minutes		30 minutes		40 minutes	
		SCORE	SCAN		SCAN	SCORE	SCAN	SCORE	SCAN
Mean values*		33.74	39.58	35.99	44.82	29.85	34.18	33.13	40.42
Difference SCAN - SCORE**			5.84		8.83		4.33		7.29
Gains in 10 minutes more***	SCORE		2.25				3.28		
	SCAN			5.24				6.24	

* There is a general decrease in the average number correct (SCORE) and the number attempted (SCAN) at retest.

** The superiority of the number of items attempted (SCAN) over the number of items answered correctly (SCORE) also declines at both 30 and 40 minutes (4.33<5.84) and (7.29<8.33) at retest.

*** There is an increase in both no of items answered correctly (3.28>2.25) and no of items attempted (6.24>5.24) at retest and the rate of gain increases (3.28/6.24 = 0.52% > 2.25/5.24 = 0.43%)

Once again, these figures vary with year of initial testing. In 1970, the correlation is near zero at both 30 and 40 minutes but reaches respectively .76 and .45 in 1971. In contrast, closer correlations (most of them higher than .65) are found between score and number of attempted items at the retest for participants of both initial years of testing.

4.3.2.1. Variation in the relationship between number of items attempted and number correct at 30 and 40 minutes.

The correlation between the number of items correctly solved (SCORE) and the number attempted (SCAN) was much higher at the end of 30 minutes (r = .59) than on completion of the test (r = .29) both at the 1970/71 testing, taken as a whole with n = 78, and again in 2002/03 [the corresponding values being 0.77 and 0.52.]. Within a same testing session, a higher correlation between number correct and number attempted (an indication of the relationship between accuracy and processing speed), means that fewer errors have been made within a given number of attempted items.

So, the higher correlations at the end of 30 minutes compared with those obtained on completion of the test seem to indicate that respondents were making relatively few errors until they encountered items that were too difficult for them. At that point they may have resorted to guessing, although that is unlikely because other works (such as that of Raven, 1981; Carpenter, Just, & Shell, 1990; Vodegel Matzen, 1991) suggest that people's answers to problems that are too difficult for them are not random but guided by incorrect hypotheses stemming from neglect of the most difficult rules governing the logic of the matrix.

Turning now to the test-retest reliability of the information on number of items attempted, we may note that, whereas there is effectively no correlation between the number attempted at the end of 40 minutes in 1971/72 and the number attempted by that time in 2002/03, the correlation at the end of 30 minutes was .36. This would seem to indicate that, as they age, many people change their strategy for dealing with more difficult problems, perhaps confirming the common finding that older people in general tend to work more slowly and carefully, check their work more, and adopt less risky strategies.

4.3.2.2.4 Gender differences in number of items attempted.

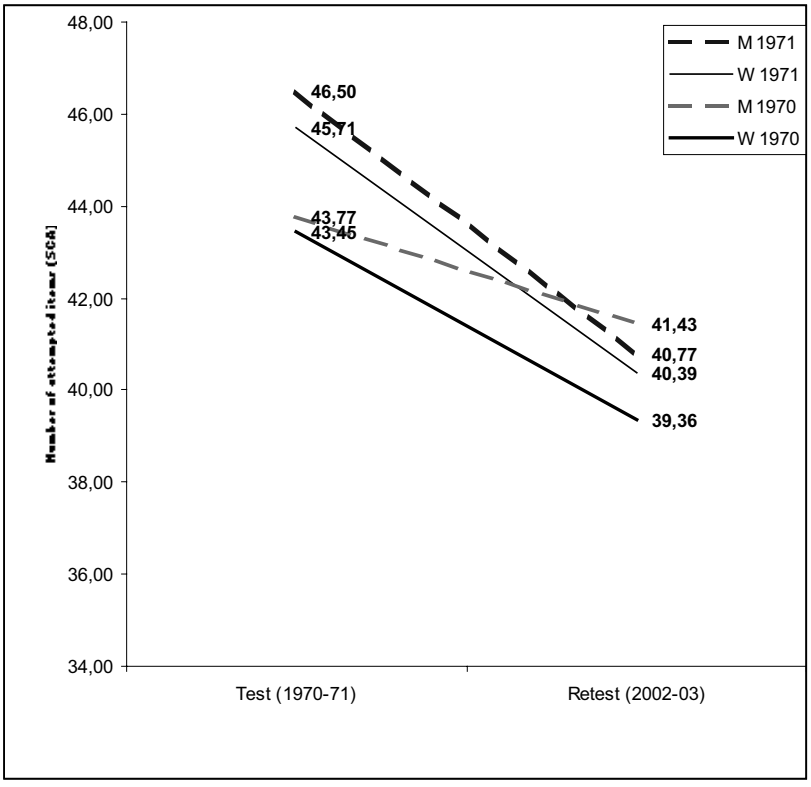
The decline in the number of items attempted for the whole test is greater for women than men – the mean decline being 4.97 items for women and 4.03 for men. The mean number items attempted by all of those

first tested in 1970 fell by 2.81 between that date and 2002/03 while the average number attempted by those first tested in 1971 fell by 5.53. The decline in the number of items attempted was least for the men first tested in 1970 (mean decline 2.34) and, unexpected as it might be, greatest for the men first tested the following year (5.74).

4.3.2.3. Speed of progress and accuracy during the 10 last minutes.

We had one last data set, the investigation of which might throw light on changes with age in people's style of work: How did what they did in the last ten minutes – i.e. between the 30 minute marker and the end of the test – change?

Figure 9.6. *Advanced Progressive Matrices, Set II, 48 Item Version, 40 Minute Time Limit*
Number of Items Attempted for Men and Women from Test (1970-71) to Retest (2002-01)
(n = 99 - 39 women / 60 men)



Within the same testing session

Compared with the number of items they had tackled per minute during the first 30 minutes of a testing session, the average respondent tackled less items and got less of those items right during the final 10 minutes of the 40 minute testing session, both at the initial test and the retest. The proportion SCORE/SCANNED is significantly lower for the last 10 minutes: more time was spent on each of these more difficult items with little gain in the number correct.

Between the two testing sessions

During the last ten minutes, our respondents not only explored, on average, one item more at their retest than they had done in the initial testing, but they were also able to solve one more item. Of course, they were working on slightly easier items and could be expected to progress more quickly than they had done 30 years before.

A more detailed study of the items that were attempted and successfully solved during the last ten minutes reveals that our respondents employed more economical strategies at the retest than they had done previously. During this time they increased their scores by, on average, only 2.25 points (out of 5.24 attempted items) at the initial testing but by 3.28 points (out of 6.24 items attempted during the last 10 minutes) at the retest, in other words the rates of success were respectively 43% and 52%. In addition to an explanation in terms of mastering easier items (the gain in score), it may be that the older respondents had developed more fruitful strategies toward the end of the test.

5. Analysis of the Global Efficiency at the Retest

As we have seen, all groups of respondents had, at the retest, on average, attempted fewer items at the end of both 30 and 40 minutes.

Nevertheless, again both at 30 minutes and at the end of the test, they solved a higher proportion of the items they attempted. In other words, they function more economically in that their restricted speed is compensated by a higher precision.

In addition, during the last ten minutes, they explore more (but easier) items than they did in the last ten minutes of the initial test and they solve a higher proportion of those easier items – just as they did during the whole retest. There is one exception to this general trend and it concerns the group of females first tested in 1971.

So, they work more slowly and carefully – but their final scores on average (and there are many exceptions to the general statement) never catch up with what they did earlier.

Conclusion

Perhaps the most important conclusion to be drawn from this study is the virtual impossibility of making meaningful generalisations. The results obtained varied dramatically from sub-population to sub-population and from individual to individual.

Only about 25% of the variance in full-time scores is explained by the variation in initial scores. Thus the rank-order of respondents' scores was very different at retest than it was at the beginning. Many people increased their scores dramatically, but similar decreases were slightly more common. The scores were more stable at the end of 30 minutes when respondents were working on the easier items and the scores obtained therefore approximated more closely to those that would have been obtained on a speed test than a power test. As they were called upon to exercise more creativity and persistence in order to solve the more difficult problems the rank order of respondents changed more. This could be due, as researchers such as Kohn and Schooler (1978, 1982), Jaques (1976), Lempert et al. (1990), and Naoi & Schooler (1990) have claimed, to variation in the demands made upon them to deal with complex problems in a work or family situation or it could be due to variation in motivation. There is no way of telling from the data available to us.

There was, however, a fairly large and intriguing difference between the results obtained from the group first tested in 1970 and the group first tested in 1971. The test-retest correlations for the first group were 0.33 and 0.42 0.36 at 30 and 40 minutes respectively for the first group and 0.75 and 0.74 0.62 for the second group. Attempts to elucidate the reasons for the difference (not reported here) were unsuccessful.

On average the scores fell somewhat (by, on average, 2.5 items) but, as noted, there was enormous individual variation with almost as many respondents showing an increase in scores as showing a decline … with some of the changes (in both directions) amounting to 9 or 10 items or more.

The next question - the question the study was undertaken to answer – is how much of the secular increase in the scores of respondents of

the same age that Thorndike and Garfinkel, Flynn, and Raven have so clearly demonstrated accounts for the apparent decline in scores with age that the data collected in cross-sectional studies have so often been interpreted to imply.

Despite considerable individual variation in the increases or decreases in scores with increasing age it is clear that it is simply not true that the reasoning – or "meaning making" – capacity of most individuals can be expected to decline at the rate previously anticipated from approximately 20 to 50 years of age. Despite the unexpected difficulties encountered when trying to interpret changes in the norms, it is obvious that the average decline in scores between these ages is trivial when compared with the secular increase in scores with date of birth that has occurred among people of similar levels of ability over the same period. Furthermore, it would seem clear from the data presented that the variation in the individual increases or decreases in scores over these 30 years is more likely to be accounted for by such things as whether (as researchers like those mentioned above have claimed) people have found themselves in situations in which they have to investigate complex problems, rather than biological processes.

A fairly extensive trawl for material which would cast light on changes in strategies with age, and especially on the question of whether it takes older people longer to achieve given levels of accuracy (without changing the *difficulty* levels of the most difficult problems they are able to solve correctly), failed to yield many conclusive results. This was due to the unreliability of the indices of number of problems attempted, the variation in the results from one sub-population to another, and the effects of introducing timing into the administration of a "power" test constructed according to the principles of Item Response Theory. The discussion may thus play an important role in alerting researchers to the difficulties that are likely to confront them when designing and interpreting studies in the area.

Notes

9.1. It follows from this that scores on the APM cannot be expected to produce a Gaussian distribution and it is obvious from Figure 9.1 that the within birth-cohort (age group) scores on the SPM are not Gaussianly distributed either. These observations render all attempts (such as Flynn's) to reduce the analysis to means and standardisations or to apply conventional statistical analyses – such as significance testing and regression analyses – inappropriate.

References

Bayley N. (1970). Development of mental abilities. In P. M. Mussen (Ed.), *Carmichael's Manual of Child Psychology.* New York: Wiley.

Bouvier, U. (1969). *Evolution des cotes à quelques Tests.* Belgium: Centre de Recherches, Forces Armées Belges.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97(3)*, 404-431.

Deary, I., Whalley, L. J., Lemmon, H., Starr, J. S., & Crawford, J. R. (2000). The stability of individual differences in mental ability from chidhood to old age: follow-up of the 1932 Scottish Mental Survey. *Intelligence*, *28*, 49-55.

Deary, I. (2000). Looking down on human intelligence: From psychometrics to the brain, *Oxford Psychology Series, 34.* Oxford: Oxford University Press.

Florquin, F. (1964). Les "P.M. 47" (PMA 1 ET 2) de J.C.Raven au niveau des classes terminales du cycle secondaire. *Revue belge de Psychologie et de Pédagogie*, XXVI, 108.

Flynn, J. R. (1984). IQ gains and the Binet decrements. *Journal of Educational Measurement, 21*, 283-290.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29-51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171-191.

Foulds, G. A., & Raven, J. C. (1948). Normal changes in the mental abilities of adults as age advances. *Journal of Mental Science, XCIV(394)*, January, 133-142.

Garfinkel, R., & Thorndike, R. L. (1976). Binet item difficulty: Then and now. *Child Development, 47*, 959-965.

Huteau, M. (2002). Psychologie Différentielle, cours et exercices (2nd Ed.). Paris: Dunod.

Jaques, E. (1976). *A General Theory of Bureaucracy*. London: Heinemann.

Kohn, M. L., & Schooler, C. (1978). The reciprocal effects of the substantive complexity of work and intellectual flexibility: A longitudinal assessment. *American Journal of Sociology, 84*, 24-52.

Kohn, M. L., & Schooler, C. (1982). Job conditions and personality: A longitudinal assessment of their reciprocal effects. *American Journal of Sociology, 87*, 1257-86.

Lempert, W., Hoff, E. H., & Lappe, L. (1990). *Occupational Biography and Personality Development: A Longitudinal Study of Skilled Industrial Workers*. Berlin: Max Planck Institute for Human Development and Education.

McCall, R. B., Appelbaum, M. L., Hogarty, P. S. (1973). Development change in mental performance. *Monographs of the Society for Research in Child development*, *38*, 150.

Mc Call, R. B. (1979). The development of intellectual functioning of later I.Q. In J. Osofsky (Ed.), *Handbook of Infant Development*. New York: Wiley.

Naoi, M., & Schooler, C. (1990). Psychological consequences of occupational conditions among Japanese wives. *Social Psychology Quarterly, 53*, 100-116.

Owens, W. A. (1959). Is age kinder to the initially more able? *Journal of Gerontology, 14*, 334-337.

Owens, W. A. (1966). Age and mental abilities: A second adult follow-up. *Journal of Educational Psychology*, *57*, 311-325.

Plassman, B. L. et al. (1995). Intelligence and education as predictors of cognitive state in late life: a 50-year follow-up. *Neurology, 45,* 1446-1450.

Prieler, J. A., & Raven, J. (10/20/02). The Measurement of Change in Groups and Individuals, with Particular Reference to the Value of Gain Scores: A New IRT-Based Methodology for the Assessment of Treatment Effects and Utilizing Gain Scores. *WebPsychEmpiricis*t http://www.wpe.info/papers_table.html

Pushkar Gold, D., Andres, D., Etezadi, J., Arbuckle, T., Schwartzman, A., & Chaikelson, J. (1995). Structural equation model of intellectual change and continuity and predictors of intelligence in older men. *Psychology and Aging, 10,* 294-303.

Raven, J. (1981). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No. 1: The 1979 British Standardisation of the Standard Progressive Matrices and Mill Hill Vocabulary Scales, Together With Comparative Data From Earlier Studies in the UK, US, Canada, Germany and Ireland*. San Antonio, TX: Harcourt Assessment.

Raven, J. (1998, updated 2003). The "decline" of eductive ability in adulthood. In J. Raven, J. C. Raven, & Court, *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1: General Overview*. San Antonio, TX: Harcourt Assessment.

Raven, J., (2000), The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive* Psychology, 41, 1-48.

Raven, J. C. (1948). The comparative assessment of intellectual ability. *British Journal of Psychology, 39*, 12-19.

Raven, J. C. (c 1950). *Progressive Matrices (1947): Plan and Use of the Scale with The Report of An Experimental Survey Carried Out by G. A. Foulds.* London: H. K. Lewis.

Raven, J. C. (1965). *Advanced Progressive Matrices: Plan and Use of the Scale with a Report of Experimental Work carried out by G. A. Foulds and A .R. Forbes.* London: H. K. Lewis. nb this is NOT the same reference as that cited on the line above.

Schaie, K. W. (Ed.). (1983). *Longitudinal Studies of Adult Psychological Development*. New York: Guilford Press.

Schaie, K. W., & Willis, S. L. (1986). *Adult Development and Ageing* (2nd edition). Boston: Little Brown

Schutzenberger, A.A. (1963). Matrix 1947. *Manuel d'Instructions et Etalonnages.* Paris: Editions Scientifiques et Psychotechniques.

Schwartzman, A. E., Gold, D., Andres, D., Arbuckle, T. Y., & Chaikleson, J. (1987). Stability of intelligence: A 40 year follow up. *Canadian Journal of* Psychology, 41, 244-256.

Thorndike, R. L. (1975). *Mr. Binet's Test 70 Years Later*. Presidential Address to the American Educational Research Association.

Thorndike, R. L. (1977). Causation of Binet IQ decrements. *Journal of Educational Measurement, 14*, 197-202.

Van Dam, F. (1976). Les "Advanced Progressive Matrices I & II" de J.C.Raven au niveau des premières candidatures en sciences, *Revue belge de Psychologie et de Pédagogie,* XXXVII, 155, 65-76.

Vernon, P. E. (1947). The variations of intelligence with occupation, age and locality. *British Journal of Psychology (Statistical Section), 1(1)*, 52-63.

Vodegel-Matzen, L. B. L. (1994). *Performance on Raven's Progressive Matrices*. Ph.D. Thesis, University of Amsterdam.

Chapter 10

The Standardisation of all the Main *Raven Progressive Matrices* Tests in Slovenia*

Dusica Boben
Center za psihodiagnosticna sredstva Ljubljana

Abstract

The *Standard Progressive Matrices* (and possibly other RPM tests) were used in the former Yugoslavia (of which Slovenia formed a part) since at least the early 1960s. More recently, i.e. since 1999, the CPM, SPM, SPM *Plus*, and APM have been standardised in Slovenia. In each case, new item analyses were carried out and the tests shown to work in similar ways to other countries and, within Slovenia, for different ability and socio-economic groups. As far as comparative norms are concerned, it seems that, allowing for the universal increase in norms over time, the Slovenian norms are similar to those obtained in other European countries.

Introduction

Let me first briefly introduce Slovenia: its history, population, school system and test use.

Slovenia and its Population

Slovenia is a small Central European country with a rich history. It gained its independence in 1991. Before that, the Slovene people lived in different countries, political regimes and cultural circles. Until World War I, Slovenians lived in the Austro-Hungarian Empire. After World

* A version of this paper containing more details on the results of the item analyses and other topics is available on the Web Psych Empiricist: http://WPE.info

War I, they united with other Slavic peoples in the Kingdom of Serbs, Croats and Slovenes. After World War II, most of the Slovenians lived in Yugoslavia. Nowadays, some Slovenians still live in Austria, Italy, Hungary, and Croatia.

Despite its small size, Slovenia's geography is very diverse. Its 20,000 square kilometres cover Alps, the Pannonic Plain, Karst, and the Adriatic Coast, as well as several lakes and thermal wells. The country has two million inhabitants and is divided into 12 regions that differ in size (see Table 10.1) and the living standard of their inhabitants. In general we could say that the west is more developed than the east. Approximately 25% is rural. 600,000 people live in the capital, Ljubljana, and its surroundings. 87% of Slovenia's inhabitants are Slovene. Approximately 12% have Slovene citizenship but belong to other ethnic groups. The proportions vary with region. The most important ethnic groups are Croats (2.8%), Serbs (2.4%), Muslims (1.4%), Hungarians (0.4%), Macedonians (0.2%), Montenegrins (0.2%), Albanians (0.2%), Italians (0.2%), Romanians (0.1%). Nationals from the former Yugoslavia live in all regions. 99% of Hungarians live in the North-Eastern Pomurska region, 90% of Italians live in the South-Western Obalno-kraska region and 94% of Romanians live in the Pomurska, Podravska, Dolenjska and Osrednjeslovenska regions. Italians and Hungarians are recognised national minorities in these respective regions, meaning that their language is also an official one,

Table 10.1. Slovenian Population Distribution

	Region	Total Population	% Population	Population Aged 5-19	% Aged 5-19
1	Pomurska	129,946	6.6	17,494	6.2
2	Podravska	320,800	16.3	43,240	15.3
3	Koroska	73,789	3.8	11,151	3.9
4	Savinjska	255,278	13.0	37,996	13.4
5	Zasavska	47,356	2.4	6,550	2.3
6	Spodnjeposavska	72,260	3.7	10,588	3.7
7	Dolenjska	95,066	4.8	15,616	5.5
8	Osrednjeslovenska	501,900	25.5	73,401	25.9
9	Gorenjska	191,688	9.8	29,398	10.4
10	Notranjsko-kraska	49,927	2.5	7,168	2.5
11	Goriska	128,124	6.5	17,392	6.1
12	Obalno-kraska	99,854	5.1	13,391	4.7
	Total	1,965,988	100.0	283,385	100.0

including use in education. The majority of the inhabitants of Slovenia are Roman Catholics Statisticne Informacije (Rapid Reports) 1992, 1997.

The number of inhabitants of Slovenia is constantly slightly decreasing. There are approximately 22,000 -- 27,000 children in one generation. The average proportion between the sexes is 51:49 (males-females).

School System

In 1990 Slovenia began to reform its school system. After the reform, obligatory schooling starts at six years of age (formerly seven) and lasts nine years. Before that, the so called "preschool", which was not obligatory, had been organised for six-year old children, followed by compulsory eight-year primary school. Regarding the new school system, one of the biggest changes as to the content and organisational level is working with children with special needs.

After completing compulsory education at a primary school, adolescents have the opportunity to continue their education at one of the secondary schools. These basically belong to three categories: general secondary schools (*gimnazija*, roughly equivalent to the German *Gymnasium*, that lasts four years and prepares their students for university study; professional secondary schools, that last four years; and vocational schools that can last from 2 ½ years to 4 years. Secondary schooling is not compulsory. About 98% of adolescents embark on it but only two thirds finish it successfully.

Tests and Testing

Applied psychology and psychological testing in Slovenia dates back to the period before World War II. Vlado Schmidt, referred to as the pioneer of applied psychology (Pečjak, 1983), was also the first to work on adaptations of group psychological tests (Lapajne, 1997). An independent Chair of Psychology within the Faculty of Arts of the University of Ljubljana was only founded in 1950. After that, systematic work on psychological tests began. In 1977, Center za psihodiagnosticna sredstva was founded. At that time it was one of the units of the state Agency for Work Productivity. Via its centres in different parts of the country, this Agency developed psychological tests in the Serbian and Croat languages as well as Slovenian.

After Slovenia gained independence in 1991 and changed its political and economic system, work continued and the need for internationally recognised work of high quality grew. The circumstances in which

psychologists work changed as well with the coming of liberal economy. Test developers had greater accessibility to foreign tests. Although demands increased financial means remained scarce.

Historical Sketch of the Usage of Raven's Progressive Matrices (RPM) in Slovenia

According to MacIntosh (1998), the RPM are one of the most widely used tests of general cognitive ability. In the course of an international survey, Oakland (1995) found that the RPM are the second most widely used psychological tests in the world. It is probably unnecessary to underline the presence of RPM in practical work as well as in basic research. The reasons for this are numerous: Simple, individual or group administration, and non-verbal items that can be used regardless of language and culture. Numerous references cited in the RPM *Manual* prove that RPM tests are indeed present in all of the five continents. Immediately after (re)publication all three forms of the RPM became the best selling tests in Slovenia.

The *Standard Progressive Matrices* (SPM) has been used in Slovenia since the beginning of the sixties, at which time Slovenia formed part of Yugoslavia. The first attempt to provide a manual for the test dates back to 1966, when the state agency for the productivity of work (Zavod SRS za produktivnost dela) published the so-called "Test information", which presented data from the 1957 edition of the British manual and data from Vito Ahtik's 1955 research on the test (that took place in Ljubljana). At that time, the SPM was mainly used in Slovenia to normalise primary school classes. After several years of use, psychologists began slowly to refrain from using the SPM for this purpose, as the results failed to express normal distribution. It was believed that the increase in scores was due to too great famiarity with the SPM items, although we now know that it was due to the so-called "Flynn's effect". Other RPM forms were not available.

In 1996, Center za psihodiagnosticna sredstva signed an adaptation agreement for all of Raven's tests (matrices and verbal scales) with J. C. Raven Ltd. We began to standardize the three classical forms: CPM, SPM and APM for pupils and young people aged 6 to 19. Simultaneously we gathered data for students and adults but, since these samples were not representative of the general population the results will not be summarised here.

The programme to standardise the classical form of the CPM, the classical form of the SPM and the APM II began in 1997. The first four

volumes of the Slovenian translation of the Manual were published in 1999 (Raven, Raven, & Court, 1999a,b,c&d). The project was carried out at the Center za psihodiagnosticna sredstva under the leadership of Dušica Boben, in co-operation with several Slovene psychologists and students of psychology, and under the supervision and with the help of John and Jean Raven. The psychometric data from the standardisation were published in a supplement to the SPM *Manual*: "The Slovene Standardisation of RPM" in 2003.

However, as elsewhere in the world, it became apparent that norms for the SPM-*Plus* were necessary. The main reason was that a reform of the educational system included a recommendation that the RPM be used to identify talented children among primary school students. As the introduction of the nine-year primary school system was a gradual one, the identification of talented students took place both in the fourth and in the eighth grade of primary school. In practice, however, the SPM proved to be too easy for students of the eighth grade. On the other hand, the APM proved to be too difficult and lacking adequate norms for younger students. In 2004, the decision to standardise the SPM-*Plus* for children between age 10.5 and 14.5 was made. This standardisation was completed in 2005 and resulted in the norms published in a supplement to the SPM-*Plus Manual* ("Slovenske norme za mladostnike v primerjavi z drugimi normami" -- Slovenian norms for adolescents in comparison to other norms).

In 2006, as part of dissertation research carried out by two psychology students (de Reggi, 2007, Klopčič, 2007), SPM-*Plus* norms for adolescents aged 14 to 17 and adults aged 38 to 53 were collected. The results for adolescents were analysed using item response theory in addition to the classical test theory.

We also started to adapt the Mill Hill Vocabulary Scales. Two pilot studies were performed (Plut, 2003, Žalik, 2003) using a sample of primary school students. This adaptation has not yet been completed, and is not, therefore, included in this presentation.

In the remainder of this chapter, the results of the standardisation of the CPM, SPM and APM in 1998 will be presented, together with those from the standardisation of the SPM-*Plus* for children and adolescents in 2005 and 2006. These data will be compared with results obtained in other countries. Most of the analyses reported below have been conducted according to classical test theory.

CPM Standardisation, 1998

Sampling and the Sample

Population of primary school students

The samples for the standardisation of all three Classic versions of the tests (CPM, SPM and APM) were drawn at the same time following a stratified random sampling procedure. First, schools were randomly selected from a list of all schools (Research results, 1996). The number of schools was set according to regions and proportionally to the number of children in a certain region. Altogether, we chose 42 schools: 29 primary and 13 secondary schools (vocational, professional, and general). The schools were requested to take part in the project and if one of them refused, another was selected from the same region using the same key. At some schools we collected data for all three tests. At others, only data for one or two tests were collected. In the CPM sample we also included children from special schools. In the CPM and the SPM samples we included one primary and one secondary bilingual school.

We assumed that the regional sampling system would capture children from more and less developed parts of the country, and children of different social-economic status. No data regarding the education and ethnic origins was collected. Altogether, 49% of data collected was from Eastern Slovenia (regions 1 to 7), 28% from Central Slovenia (including Ljubljana), and 23% from Western Slovenia (regions 9 to 12). The percentage of data collected in the different regions corresponds to the proportion of children in those regions.

The data collection coordinators at individual schools were requested to select one class from each grade. Parents of the selected children were sent a written presentation of the project and a request for co-operation. Very few requests were refused. Testing took place in 1998 and was performed by school psychologists, psychologists of the Center za psihodiagnosticna sredstva, and assisted by several final year psychology students. An educational event was organised for the test administrators, where the project of standardisation was presented, and test administrators were trained to administer the test (testing instructions, conditions etc…). Testing was performed as group testing, it took place at the schools in time of lectures and without time limitation. Data was processed using STATISTICA software (StatSoft, 1999). The norms were calculated by John and Jean Raven.

The CPM sample included students from 1st to 6th grade drawn from 23 schools in 20 different cities and towns of various sizes from all of the regions. In the end, 1,230 children aged 6 ½ to 14 were tested. This amounts to 0.85% of the population of this age. For the calculation of one year age norms we considered the results of 1,199 children (Table 10.2). 53% of them were male; 88% of them came from regular primary schools.

Population of pre-school children

The sample of pre-school children was planned within the framework of research towards a PhD. thesis bearing the title *Development of phonological conscience at pre-school children* (Jerman, 2000). 541 children aged 6 to 7 ½ were included in the sample (Table 10.3). All of them were involved in pre-school programmes in 29 kindergartens in 29 different towns in Slovenia. The sample represents 0.95% of the population of Slovene children of this age. 48.7% of children in the sample were male. The children were tested individually and without time limitation. The testing was performed by 32 psychology students and psychologists that had previously been trained for this type of testing.

Item analysis

The (conventional) difficulty indexes of the CPM items largely correspond to those established in the British studies. The small differences that did exist are discussed in the Web Psych Empiricist: http://WPE.info version of this article, where the results of a distractor analysis are also presented.

Table 10.2. *Coloured Progressive Matrices*
1998 Slovenian Sample (Primary School)

Age in Years	Age Years (Months)	Male	Female	Total
7	6(6)-7(5)	27	28	55
8	7(6)-8(5)	174	160	334
9	8(6)-9(5)	120	83	203
10	9(6)-10(5)	107	94	201
11	10(6)-11(5)	85	94	179
12	11(6)-12(5)	64	63	127
13	12(6)-13(5)	49	51	100
Total		626	573	1,199

Table 10.3. *Coloured Progressive Matrices*
1998 Slovenian Pre-school Sample

Age	Age Years (Months)	*n*
6	5(9) - 6(2)	113
6½	6(3) - 6(8)	234
7	6(9) - 7(2)	178
7½	7(3) - 7(8)	16
Total		541

Internal consistency

The internal consistency of the CPM assessed from the Slovenian sample of primary school students described above was 0.89 (Cronbach alpha), or 0.91 (split-half). These figures are similar to those found in other countries and continents (Raven et al., 1999a). The average correlation between the items was 0.20.

The internal consistency improved with age, rising from 0.86 among seven year olds to 0.92 among thirteen-year olds. These findings confirm those from other studies (Raven, Raven, & Court, 1999a).

The internal consistency assessed from the sample of preschool children was 0.90 or 0.89 (standardised Cronbach alpha coefficient).

Gender differences

There were no significant gender differences within age groups ($F = 0.11$, $p = 0.74$) or any gender-age interaction ($F = 0.73$, $p = 0.62$).

Speed of work

In the course of testing, the time needed to complete the test was recorded. The shortest was four minutes and the longest 33 minutes. On the average, children aged 7--13 completed the CPM in 10 minutes (with a standard deviation of 4.3 minutes). The older children required less time and there was less variability between them. 7-year olds needed 13 ½ minutes on the average (standard deviation = 5.2), whereas 11-year olds needed 8.7 minutes (standard deviation = 5.2).

Summary of Results

The means, standard deviations, skewness and kurtosis are presented, by age group, in Table 10.4. All distributions are left asymmetric, and, as expected, the most asymmetric distribution is in the group of oldest

Table 10.4. *Coloured Progressive Matrices*
Mean (M), Standard Deviation (SD), Skewness And Kurtosis for Different Age Groups

Age	Age Years (Months)	*n*	M	SD	Skewness	Kurtosis
Pre-School						
6	5(9)-6(2)	113	22.8	6.3	-0.79	0.20
6½	6(3)-6(8)	234	22.8	6.7	-0.81	-0.07
7	6(9)-7(2)	178	24.6	7.1	-0.64	-0.22
Primary School						
7	6(6)-7(5)	54	25.2	5.8	-1.04	1,44
8	7(6)-8(5)	334	25.9	5.8	-0.83	0,68
9	8(6)-9(5)	203	27.5	5.9	-0.88	0,26
10	9(6)-10(5)	200	28.7	5.9	-1.28	1,23
11	10(6)-11(5)	179	30.2	4.8	-1.32	1,83
12-13	11(6)-13(5)	227	31.1	5.8	-1.77	2.68

children. None of the distributions is explicitly bimodal, as are some found in the literature (e.g. Raven, 1981). The detailed normative data will be presented and discussed later.

SPM Standardisation, 1998

Sampling and the Sample

As described in the previous section, the SPM sample was drawn at the same time as the CPM one. The SPM sample included secondary pupils as well as primary school pupils.

The expression "secondary school" in this context refers to all three categories of secondary school: vocational, professional, and general ("*gimnazije").* Thus it included students of educational programmes lasting three years, preparing them for vocations such as hairdressers, painters, car mechanics etc., students enrolled in educational programmes lasting four or five years and awarding professional qualifications such as mechanical technician, chemistry technician, construction technician, and students enrolled in general 'gimnazija' programmes.

Overall, it covered students from the 1st to the 8th year at 10 primary schools and students of the first and second year at 14 secondary schools -- altogether 1,556 children and adolescents aged 7 ½ --18 years (Table 10.5). This amounts to 0.6 percent of the population. The youngest and

Table 10.5. *Standard Progressive Matrices*
1998 Slovenian Sample

Age	Age Years (Months)	Male	Female	*n*
8	7(6)-8(5)	53	46	99
9	8(6)-9(5)	71	57	128
10	9(6)-10(5)	60	55	115
11	10(6)-11(5)	73	52	125
12	11(6)-12(5)	58	65	123
13	12(6)-13(5)	55	61	116
14	13(6)-14(5)	65	67	132
15	14(6)-15(5)	70	74	144
16	15(6)-16(5)	146	137	283
17	16(6)-17(5)	115	96	211
18	17(6)-18(5)	51	29	80
Total		817	739	1,556

oldest students were somewhat under represented. 53% of the sample were male students.

Item analysis

In general, the difficulty indexes for the SPM items match the original British ones very well. The differences are discussed in the Web Psych Empiricist version of this article. The correlations between the item difficulties established separately within age group ranged from 0.76 (between 8- and 18-year olds) to 0.99 between two "neighbouring" age groups. These are comparable to the information from Great Britain (Raven, 1981). As with the CPM, a distractor analysis was also carried out and the results reported in the WPE version of this article.

Internal consistency

Both Cronbach alpha and split-half internal consistency indices were .95. The average correlation between the items was 0.22. Cronbach alphas calculated within one year age groups varied only slightly around the overall figure. Hence, there are no great differences between older and younger children, like the ones they discovered in Great Britain (Raven, 1981).

Gender and age differences

Anova shows that subgroups differ in statistically significant ways in relation to sex ($F = 13.13$, $p = 0.00$) and age group (one year intervals) from 8 to 18 years ($F = 76.48$, $p = 0.00$), but not regarding the interaction between them ($F = 0.65$, $p = 0.77$). A more detailed analysis shows that sex differences occur only in the older age groups. T-tests revealed statistically significant differences for age groups of 16-year olds ($p = 0.02$), 17-year olds ($p = 0.01$) and 18-year olds ($p = 0.04$). Nevertheless, statistically significant differences regarding sex were not confirmed by Tukey's HSD test for individual subgroups, separated by age. The bar diagrams in Figure 10.1 show the frequency distribution of raw scores for males and females. The females scored slightly higher in all age groups. In Great Britain, higher results were only achieved by girls older than 12 (Raven, 1981). Perhaps, in the Slovenian situation, the larger differences between the sexes at age of 16 to 18 could be explained by motivation.

Raw scores distributions and descriptive statistics

The *average time* required to complete the test was 25 minutes (SD = 6.7) The minimum was 8 minutes and the maximum 33 minutes.

Means, standard deviations, skewness and kurtosis for males and females combined for individual age groups from age 8--18 are presented in Table 10.6.

All distributions, except the one for the 8-year olds' age group, are left asymmetric. The distributions for the 8 and 9-year old groups are bimodal, as in the British 1979 standardisation (Raven, 1981). For ages 9, 10 ½ and 12, bimodality also appears as in the Irish 1972 standardisation (Raven, 1981). As peaks appear at different values of the raw result and disappear at larger sub-samples, Raven (1981) suggests that they may be due to variance in the quality of the samples that is inevitably associated with random sampling. Other explanations are also possible, for example the adoption of different strategies for solving the problems (Lake, in Raven, 1981).

Figure 10.1. *Standard Progressive Matrices*
Distribution of Scores by Gender

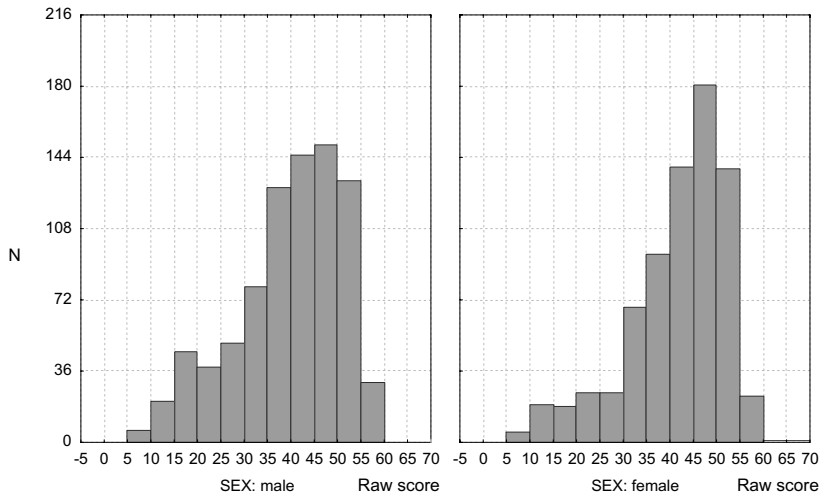


Table 10.6. *Standard Progressive Matrices*
Mean (M), Standard Deviation (SD), Skewness and Kurtosis for Different Age Groups

Age	Age Years (Months)	*n*	M	SD	Skewness	Kurtosis
8	7(6)-8(5)	99	24.8	9.2	0.17	-0.77
9	8(6)-9(5)	128	30.3	10.8	-0.36	-1.02
10	9(6)-10(5)	115	36.0	9.2	-0.29	-0.45
11	10(6)-11(5)	125	38.4	9.8	-1.00	0.60
12	11(6)-12(5)	123	42.1	7.8	-0.91	0.78
13	12(6)-13(5)	116	42.9	9.1	-1.35	2.28
14	13(6)-14(5)	132	42.6	9.0	-1.21	1.90
15	14(6)-15(5)	144	45.8	8.9	-1.40	2.35
16	15(6)-16(5)	283	45.5	8.9	-1.46	3.16
17	16(6)-17(5)	211	46.0	9.1	-1.40	1.82
18	17(6)-18(5)	80	46.4	7.7	-0.88	0.46

SPM *Plus* Standardisation in 2005 and 2006

Sampling and the Sample

Adolescents aged 12 to 14

As discussed in the introduction to this article, we decided to generate norms for SPM *Plus* (SPM *Plus*) for adolescents aged 12 to 14 (i.e. students of the 6th to 8th grade of primary school). The sampling process was similar to that employed to generate the 1998 sample discussed above. First, the number of schools from a certain region was set, depending on the size of the region. Second, schools were picked from the list of all schools that had agreed to co-operate with us in the project of standardisation. In the selection of schools, the distinction between schools from smaller towns (population under 6,000) and larger towns was respected, as was the proportion of such towns in the Slovenia. In every selected school, we designated a grade (year) and a class in that grade, from which data were to be collected. Parents' and school management's co-operation was requested. Testing took place in groups, in the morning, and without time limitation. It was performed in 2005 by school resident psychologists and psychologists of Center za psihodiagnosticna sredstva, all of whom had had previous experience with RPM testing.

1,079 students were included in the final sample, aged 10 to 15 ½ (Table 10.7), but only the results of children aged 10 ½ --14 ½ were

Table 10.7. *Standard Progressive Matrices Plus*
2005 and 2006 Slovenian Samples, Age, and Sex

Age	Age Years (Months)	Male	Female	*n*
Sample 2005				
11	10(6)-11(5)	33	45	78
12	11(6)-12(5)	169	152	321
13	12(6)-13(5)	183	186	369
14	13(6)-14(5)	139	149	288
Total		524	532	1056
Sample 2006				
14	13(6)-14(5)	37	38	75
15	14(6)-15(5)	91	103	194
16	15(6)-16(5)	106	103	211
17	16(6)-17(5)	65	65	130
Missing sex data				2
Total		299	309	610

included in our analyses. There were 1,056 such students. 50% of students included in the sample were female. 29% came from schools in towns with less than 6,000 inhabitants. Both these figures correspond well to the proportion in the population. Altogether, 25 primary schools took part in the project, 18-19 for each sub-sample.

15- and 17-year old adolescents

In 2006, Center za psihodiagnosticna sredstva agreed to co-operate in a project leading towards a graduation thesis entitled "*Psychometric characteristics of Raven's SPM-Plus regarding Slovene adolescents*" (de Reggi, 2007). The target population consisted of students of the ninth grade of primary school and of the first and second year of secondary schools. Schools were selected, as in the other research projects, according to the proportion of the population of the individual region. Regions with smaller populations were joined to neighbouring regions so that they are commonly represented in the sample. Ten primary schools and 10 secondary schools (three general secondary schools, five professional secondary schools, and two vocational schools) from all over the country were included in the sample. Schools, within the statistical region and educational programme categories, were selected randomly using a telephone directory. Principals from two schools declined co-operation but very few parents did so.

Testing was performed by resident school psychologists and psychologists from Center za psihodiagnosticna sredstva. There was no time limit.

The final sample included 610 adolescents, aged 14 to 17 (Table 10.7). Among them, 184 were students of the 9th grade of primary school, 225 of them were students of the first year of secondary school and 201 were students of the second year of secondary school. 49% of the students in the sample were male. There were fewer 14-year olds (n = 75), as this age group was not the target group of the project, as the norms for this age group had already been collected.

Item analyses

The correlation between the item difficulties for the SPM *Plus* (calculated in the traditional way - i.e. proportion choosing the correct answer) established separately in the first and the second samples described above was 0.998 (or 1.00 to two decimal places). These item difficulties also correspond to those published in the British *Manual.* It is important to

note, however, that, as in the UK, the item difficulties do not increase steadily within Sets. The largest deviations are around B8-B12, C4-C8, and D6-D10. There is also a large discrepancy between the most difficult items of set C and the easiest items of set D. As explained in the original *Manual* (Raven et al 2000, updated 2004), this arose from the need to merge items from different Sets in the *Parallel* version of the *Classic Standard Progressive Matrices* to make room for the more difficult items in the SPM ***Plus***. Despite these deviations from the, in some senses, ideal order of items, the net effect has, as can be seen from the graphs of item difficulties published in Raven et al (2000, updated 2004) and in the chapters reporting the results of the Romanian standardization of the SPM *Plus* in this volume, been the production of a test having an almost linear relationship between total score and item difficulty and an almost linear Test Characteristic Curve. This has enormous benefits from the point of view of avoiding misinterpretations of research and the calculation of change scores.

De Reggi re-analysed the data using a three-parameter Item Response Theory model (as operationalised in Bilog-MG software [(Du Toit, 2003)] and compared her results with the Romanian data reported elsewhere in this volume. The correlation between the conventional item difficulties (as reported above) and those derived from the IRT based procedures was 0.95, and the correlation between the IRT-based difficulty parameters from the Slovenian and Romanian sample was also 0.95. The greatest deviation can be observed at extreme values and can probably be attributed to the narrower age base of the Slovenian sample.

Distractor analysis revealed that misleading distractors (cases where adolescents chose one of the false answers more often than the correct one) are most often to be found in Set E, which is, of course, the most difficult Set. (This can be understood as the test administration instructions encourage guessing.) Similar results were found for the last items of sets C and D.

Internal consistency

Like most of the authors of other chapters in this book, we sought to assess the internal consistency of the SPM *Plus* without fully appreciating the inappropriateness of intercorrelating the items of IRT-based tests (as explained in the General Introduction to this book). The Cronbach alpha coefficient derived from our sample of 1079 was 0.81, and the standardised Cronbach Alpha 0.80. The split half coefficient was 0.83. The average correlation between items was 0.06. These figures are all

slightly lower than those for the Classic SPM for Slovene adolescents (Boben, 2003) and are probably due to the restricted range of scores in the sample (the test is, like the Classic SPM, intended for use with all age groups from 5 to 80 years of age). Nevertheless, they are relatively high compared with other tests and similar to those for the SPM *Plus* published by others (e. g. Matešić, 2000b, Dobrean et al., 2005)

Although the internal consistency varies with age group, the differences are small. It is lowest among 12-year olds (0.79), and highest among 11 and 13-year olds (0.82).

The internal consistency index (Cronbach alpha) of the SPM *Plus* for the second sample (n = 610) was 0.82, varying across age groups from 0.78 to 0.83. The standard error of measurement was 2.59 and standard error of estimate 2.34 (de Reggi, 2007).

Gender and age differences

The data from the first sample were checked for age and sex differences. Anova confirmed age group differences (F = 13.22, p = 0.00), but not differences between the sexes (F = 3.33, p = 0.07), which, considering numerous other researches, could be expected. Detailed analysis also showed that girls were better at solving the SPM *Plus* than boys in the subgroup of 11-year olds, which was the smallest group (t = -3.42, p = 0.001). This can probably be attributed to the size of the sample, which was too small and allowed different motivation of tested students to affect the results. Girls, on the average, achieved better results than boys in all of the age groups and in the entire sample. The results of the Romanian research for the entire sample were similar (Raven, Raven, & Court, 2004). Girls in Slovenia also achieve slightly higher results with the SPM (Boben, 2003).

Similar calculations were performed by de Reggi (2007). She found statistically insignificant differences between the sexes, with slightly higher average results in favour of girls aged 15--17. There were no statistically significant differences between age groups in this sample. The higher the year of tested students, the higher the average results, with one exception: 17-year olds achieved a lower average value than students who were two years younger. There were statistically significant differences between different categories of secondary schools: Students of vocational schools achieved results that were below the average values of the sample, whereas students of general secondary schools ("gimnazija") achieved results higher than the average results of the sample.

Raw scores distributions and descriptive statistics

In the first sample, the average time required to complete the SPM *Plus* was between 30 minutes (11-year olds; SD = 12.6) and 34 minutes (14-year olds; SD = 13.2). The variability in the time required was highest among the oldest adolescents, and practically the same in the case of other tested adolescents -- approximately 12 ½ minutes. The average time required to complete the test increases with the age.

In the second sample (*n*= 610), the average time taken to complete the test was very similar in all of the age groups -- approximately 26 minutes. Only variability differs between age groups, being greatest in the group of 17-year olds (9 minutes) and smallest in the group of 14-year olds (7.5 minutes).

The within-age frequency distributions for the SPM *Plus* test in the Slovenian sample are normal, as was the case for the Classic SPM for this age (11--14 years). The descriptive statistics are summarised in Table 10.8. Average SPM *Plus* scores increase with age. All distributions are somewhat left symmetric, i.e. positioned towards higher results. The Kurtosis among 11-year olds was less satisfactory, but probably due to the sample as differences between the sex groups are also greater. This was also true of 14-year olds, the achievements of whom also differ from results from the other sample. In general, we can observe that average results in the 2006 sample are lower, although one would - because of the Flynn effect - expect higher results. The results of the 17-year olds are the most surprising in the sense that the average score does not fit into the general tendency of scores to increase with age. The most probable explanation of this is lower motivation level, however another hypothesis is that socio-economic status was not explicitly controlled in the selection of the sample.

APM Standardisation, 1998

Sampling and the Sample

As mentioned earlier, the sample for the APM standardisation was drawn at the same time as that for the CPM and Classic SPM. Testing was carried out in groups, without time limitations. Only the APM II data were processed, although the APM I was used to present instructions and to check whether participants were capable of solving the test.

Table 10.8. *Standard Progressive Matrices* ***Plus*** **2005 and 2006 Slovenian Samples Mean (M), Standard Deviation (SD), Skewness and Kurtosis for Different Age Groups**

Age	Age Years (Months)	*n*	M	SD	Skewness	Kurtosis
Sample 2005						
11	10(6)-11(5)	78	31.35	6.18	-0.96	1.31
12	11(6)-12(5)	321	32.62	5.67	-0.38	0.21
13	12(6)-13(5)	371	34.08	5.89	-0.24	0.82
14	13(6)-14(5)	288	35.07	6.02	-0.56	1.24
Sample 2006						
14	13(6)-14(5)	75	33.2	5.5	-0.45	0.71
15	14(6)-15(5)	194	34.2	6.4	-0.30	1.00
16	15(6)-16(5)	211	35.3	5.9	-0.31	1.03
17	16(6)-17(5)	130	34.5	6.2	0.30	0.50

This way, the sample included students from the 6th to 8th year (grade) of nine primary schools and of all of the years of 15 secondary schools, i.e. adolescents aged 12--19. We included data collected from 1,363 adolescents, which represents 0.72% of the population of that age. 43% of the sample were male, which is slightly less than in the population. The reason for this is that the adolescents of this age, who are not enrolled in educational programmes, were not included in the sample, and this group is predominantly male.

Item Analysis

According to the difficulty indices, there are more cases than with the SPM and CPM, in which the Slovenian order of difficulty of the items differs from the original, but they are again not large and are reported in the Web Psych Empiricist version of this paper.

The frequency distribution of the difficulty indexes is bimodal, with a mean of 0.49 (SD = 0.26) and median of 0.53. There are less items of moderate difficulty and good discrimination than one would have expected in a power test. A smaller number of easy items than with CPM and SPM can also be observed.

The correlations between the difficulty indexes established separately within age groups are high (0.97--0.99).

The results of our distractor analysis are again reported in the WPE version of this paper.

Table 10.9. *Advanced Progressive Matrices*
1998 Slovenian Sample

Age	Age Year s(Months)	Male	Female	*n*
13	12(6)-13(5)	61	65	126
14	13(6)-14(5)	78	76	154
15	14(6)-15(5)	87	90	177
16	15(6)-16(5)	112	124	236
17	16(6)-17(5)	101	163	264
18	17(6)-18(5)	91	156	247
19	18(6)-19(5)	65	94	159
Total		595	768	1,363

Internal consistency

The Cronbach alpha coefficient for the APM II was 0.87 and split half 0.88. The Cronbach alpha coefficients for individual age groups ranged between 0.83 (14-year olds) and 0.88 (15-year olds). The average correlation between items was 0.15.

Gender and age differences

With sex and age (year intervals) as independent variables, Anova showed that there were differences in APM II raw score between age groups (F = 7.98, p = 0.00) but not between the sexes (F = 1.82, p = 0.18). Nor is there an interaction between age and sex.

Raw scores distributions and descriptive statistics

The average time taken to complete the test was 29 minutes (SD = 9.8); 4 minutes minimum and 66 minutes maximum. There were some differences among primary school students. Whereas the differences among students attending secondary schools are not that big and do not deviate much from the total average there are however differences in standard deviation.

The 16-year olds got unexpectedly high scores in comparison with the younger and older groups (Table 10.10). All distributions (in individual age groups) are somewhat left asymmetric, but not distinctly, as the mean remains at the half of all possible points. A distinct bimodal distribution can be observed, in both sexes, among 13-year olds. Probably, the same hypotheses could be put forward to explain them as were suggested re the SPM.

Slovene Norms in an International Context

Tables 10.11 to 10.15 present smoothed RPM age norms for different age groups. The norms show the raw score required to do better than the stated percentage of the population of each age group. Norms presented in this way have been consistently used by Raven for the past 70 years since, unlike norms presented as having a mean of 100 and a SD of 16, they (a) make no assumptions about the shapes of the within age distributions, (b) do not exaggerate the discriminative power of the test (compare an SD of 15 with SDs of 6 to 10 in the tables above), and (c) do

Table 10.10. *Advanced Progressive Matrices Set II*
Mean (M), Standard Deviation (SD), Skewness and Kurtosis for Different Age Groups

Age	Age Years (Months)	*n*	M	SD	Skewness	Kurtosis
13	12(6)-13(5)	126	15.3	6.6	-0.20	-0.93
14	13(6)-14(5)	154	16.2	5.9	-0.04	-0.91
15	14(6)-15(5)	177	17.3	7.0	-0.31	-0.79
16	15(6)-16(5)	236	19.0	6.5	-0.39	-0.40
17	16(6)-17(5)	264	17.7	6.4	-0.24	-0.36
18	17(6)-18(5)	247	18.2	6.3	-0.15	-0.08
19	18(6)-19(5)	159	19.3	6.0	-0.41	0.38

Table 10.11. *Coloured Progressive Matrices*
Smoothed 1998 Slovenian Norms (Pre-School, Individual Administration) In the Context of 1983 Dumfries Data

	Age in Years (Months)					
	6 5(9) to 6(2)		6½ 6(3) to 6(8)		7 6(9) to 7(2)	
Percentile	UK	**SL**	UK	**SL**	UK	**SL**
95	24	**33**	26	**34**	28	**35**
90	21	**31**	23	**31**	25	**34**
75	19	**26**	20	**27**	21	**30**
50	**16**	**22**	**17**	**23**	**18**	**24**
25	13	**17**	14	**17**	16	**19**
10	11	**14**	12	**13**	13	**15**
5	9	**13**	11	**11**	12	**11**
n	23	**113**	42	**234**	54	**178**

not perpetuate the images and myths associated with "IQ". All the norms presented have been smoothed to eliminate sampling error (see Raven, 2000 for a discussion of this problem.)

Table 10.11 presents CPM norms for Slovene pre-school children tested individually in the context of UK data from 1982. Table 10.12 presents CPM norms for Slovene primary school tested in groups in

Table 10.12. *Coloured Progressive Matrices*
Smoothed 1998 Slovenian Norms (Primary Schools) In the Context of 1982 Dumfries (UK) Data

	Age in Years (Months)									
	7½ 7(3) to 7(8)		8 7(9) to 8(2)		8½ 8(3) to 8(8)		9 8(9) to 9(2)		9½ 9(3) to 9(8)	
Percentile	UK	**SL**	UK	**SL**	UK	**SL**	UK	**SL**	UK	**SL**
95	31	**34**	32	**34**	33	**34**	34	**34**	35	**34**
90	28	**32**	30	**32**	32	**33**	33	**33**	33	**33**
75	23	**29**	25	**30**	27	**31**	29	**31**	31	**32**
50	**20**	**25**	**22**	**26**	**24**	**27**	**26**	**27**	**28**	**28**
25	17	**21**	18	**22**	20	**23**	22	**24**	24	**25**
10	14	**17**	15	**18**	16	**18**	17	**19**	19	**20**
5	13	**13**	14	**14**	14	**14**	15	**14**	16	**15**
n	55	**115**	44	**175**	48	**128**	52	**102**	37	**104**

Table 10.12 (continued)

	Age in Years (Months)										
	10 9(9) to 10(2)		10½ 10(3) to 10(8)		11 10(9) to 11(2)		11½ 11(3) to 11(8)		12 11(9) to 12(2)	12½ 12(3) to 12(8)	13 12(9) to 13(2)
Percentile	UK	**SL**	UK	**SL**	UK	**SL**	UK	**SL**	**SL**	**SL**	**SL**
95	35	**34**	35	**34**	35	**35**	35	**35**	**35**	**35**	**35**
90	33	**33**	34	**33**	35	**34**	35	**34**	**34**	**34**	**34**
75	32	**32**	33	**32**	33	**33**	34	**33**	**33**	**33**	**33**
50	**30**	**29**	**31**	**30**	**31**	**31**	**32**	**31**	**32**	**32**	**32**
25	25	**26**	26	**27**	28	**28**	30	**29**	**29**	**30**	**30**
10	21	**21**	22	**22**	23	**23**	25	**23**	**24**	**24**	**25**
5	17	**16**	18	**17**	20	**18**	22	**19**	**19**	**19**	**19**
n	53	**96**	49	**104**	51	**83**	55	**80**	**67**	**61**	**59**

the context of the previously mentioned UK data. Table 10.13 presents Classic SPM norms for Slovenia in the context of the 1979 British data. Table 10.14 presents the Slovenian SPM ***Plus*** norms (from the first sample) in the context of German (D), American (FB), Romanian (RO), Croatian (HR) and Polish (PL) data. Table 10.15 presents the Slovenian APM II norms in the context of British 1979 data.

By and large, the Slovene norms are remarkably similar to those obtained in other countries. The one exception seems to be that the

Table 10.13. *Standard Progressive Matrices*
1998 Slovenian Smoothed Norms In the Context of 1979 British Data

	Age in Years (Months)									
	8		9		10		11		12	
	7(9) to 8(2)	7(6) to 8(5)	8(9) to 9(2)	8(6) to 9(5)	9(9) to 10(2)	9(6) to 10(5)	10(9) to 11(2)	10(6) to 11(5)	11(9) to 12(2)	11(6) to 12(5)
Percentile	UK	**SL**	UK	**SL**	UK	**SL**	UK	**SL**	UK	**SL**
95	40	**40**	44	**46**	48	**48**	50	**50**	52	**52**
90	38	**37**	42	**42**	46	**46**	48	**48**	50	**50**
75	33	**32**	38	**39**	42	**43**	44	**45**	46	**47**
50	**25**	**23**	**33**	**31**	**38**	**36**	**40**	**40**	**41**	**42**
25	17	**18**	25	**22**	32	**28**	34	**33**	37	**37**
10	14	**12**	17	**16**	23	**20**	29	**24**	31	**30**
5	12	**10**	14	**11**	17	**14**	24	**17**	26	**23**
n	174	**99**	166	**128**	172	**115**	187	**125**	164	**123**

Table 10.13 continued

	Age in Years (Months)								
	13		14		15		16	17	18
	12(9) to 13(2)	12(6) to 13(5)	13(9) to 14(2)	13(6) to 14(5)	14(9) to 15(2)	14(6) to 15(5)	15(6) to 16(5)	16(6) to 17(5)	17(6) to 18(5)
Percentile	UK	**SL**	UK	**SL**	UK	**SL**	**SL**	**SL**	**SL**
95	54	**53**	55	**54**	57	**55**	**56**	**56**	**56**
90	52	**51**	54	**52**	55	**53**	**54**	**54**	**54**
75	49	**48**	50	**49**	51	**50**	**51**	**52**	**52**
50	**43**	**43**	**45**	**44**	**47**	**45**	**46**	**48**	**49**
25	39	**38**	42	**39**	42	**40**	**41**	**41**	**42**
10	33	**32**	36	**33**	36	**35**	**35**	**35**	**36**
5	28	**25**	30	**26**	33	**26**	**27**	**29**	**30**
n	185	**116**	196	**132**	191	**144**	**283**	**211**	**80**

Slovenian CPM norms for young children tested individually are well above those for the comparison group from the UK. The children in the UK group were, however, only tested individually if they could not cope with the answer sheets on their own. So the higher scores would seem, almost certainly, to be a product of individual testing.

Table 10.14. *Standard Progressive Matrices* ***Plus***
Comparison of Slovenian, German, Fort Bend (Texas), Romanian, Croation, and Polish Norms

	Age in Years (Months)											
	11 10(6) to 11(5)	12 11(6) to 12(5)	13 12(6) to 13(5)	14 13(6) to 14(5)	14	14 13(9) to 14(2)	14 13(9) to 14(2)	14½ 13.5 to 15.5	14½ 14(3) to 14(8)	15	15	15 14(9) to 15(2)
Percentile	**SL**	**SL**	**SL**	**SL**	**D**	**FB**	**RO**	**HR**	**RO**	**PL**	**D**	**FB**
95	40	42	43	45	43	44	41	44	42	49	45	46
90	37	40	41	42	40	41	39	42	40.1	48	43	43
75	35	37	38	39	37	39	35	38	35	44	40	40
50	**32**	**33**	**34**	**35**	**33**	**36**	**31**	**35**	**31**	39	**36**	**37**
25	28	29	30	32	29	32	24	31	25	36	32	34
10	24	25	27	28	26	30	18	27	19	33	29	31
5	21	23	24	25	24	27	15	21	15	30	27	29
n	78	321	371	288	181	24	69	295	70	98	523	24

Table 10.15. *Advanced Progressive Matrices*
Smoothed 1998 Slovenian Norms In the Context of 1979 UK Data

	Age in Years (Months)								
	12½ 12(3) to 12(8)	13 12(9) to 13(2)	13½ 13(3) to 13(8)	14 13(9) to 14(2)		14½ 14(3) to 14(8)		15 14(9) to 15(2)	
Percentile	**SL**	**SL**	**SL**	UK	**SL**	UK	**SL**	UK	**SL**
95	**23**	**23**	**24**	23	**24**	25	**25**	26	**25**
90	**22**	**22**	**23**	22	**23**	22	**24**	23	**24**
75	**19**	**19**	**20**	17	**20**	17	**21**	18	**21**
50	**15**	**15**	**16**	**12**	**16**	**13**	**17**	**14**	**17**
25	**9**	**9**	**10**	10	**10**	10	**11**	10	**11**
10	**5**	**5**	**6**	7	**6**	7	**7**	7	**7**
5	**3**	**3**	**4**	4	**4**	5	**5**	5	**5**
n	**50**	**62**	**75**	196	**66**	189	**84**	191	**71**

Table 10.15 - (continued)

		Age in Years (Months)							
		15½	16	16½	17	17½	18	18½	19
		15(3)	15(9)	16(3)	16(9)	17(3)	17(9)	18(3)	19(9)
		to	to	to	to	to	to	to	To
		15(8)	16(2)	16(8)	17(2)	17(8)	18(2)	19(8)	20(2)
Percentile	UK	**SL**	**SL**	**SL**	**SL**	**SL**	**SL**	**SL**	**SL**
95	27	**26**	**26**	**27**	**27**	**28**	**28**	**29**	**29**
90	23	**25**	**25**	**26**	**26**	**27**	**27**	**28**	**28**
75	18	**22**	**22**	**23**	**23**	**24**	**24**	**25**	**25**
50	**14**	**18**	**18**	**19**	**19**	**20**	**20**	**21**	**21**
25	10	**12**	**12**	**13**	**13**	**14**	**14**	**15**	**15**
10	7	**8**	**8**	**9**	**9**	**10**	**10**	**11**	**11**
5	5	**6**	**6**	**7**	**7**	**8**	**8**	**9**	**9**
n	171	**117**	**116**	**123**	**141**	**130**	**126**	**109**	**90**

Given the differences between the dates of testing, those familiar with the so-called "Flynn Effect" may have expected bigger differences between the Slovene norms and those collected in other countries. And, in fact, more recent data collected in some other countries (such as Switzerland and Norway) are higher than both the Slovene norms and their earlier UK counterparts (Raven et al., 1999c).

The UK SPM Manual (Raven, Raven, & Court, 1999c) presents tables (SPM3 and SPM5) to convert SPM *Plus* and APM scores to SPM scores. Doubts have sometimes been expressed about the accuracy of these conversion tables. We therefore thought it would be useful to see what would happen if we converted the SPM *Plus* and APM data we had collected from our different samples to SPM scores and compared the results.

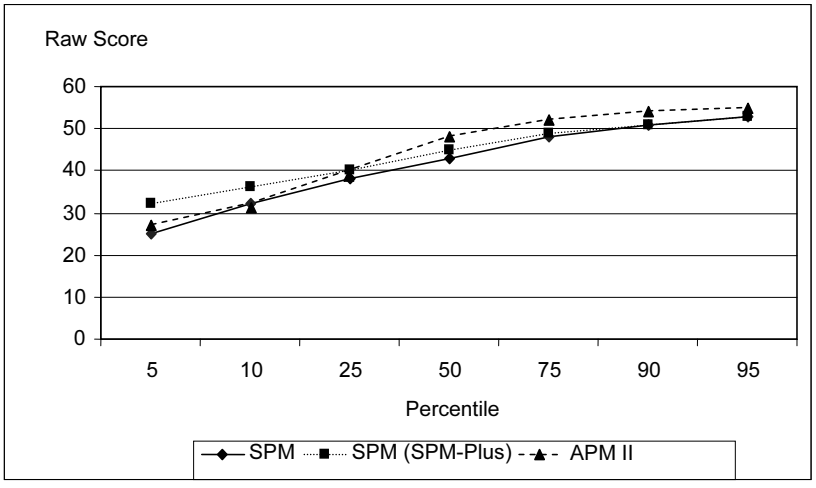
The results for 13 year olds are presented in Figure 10.2. Those for the other age groups were similar. It will be seen that the estimated raw score equivalent to the 5^{th} and 10^{th} percentiles diverges somewhat between the tests, with data from the SPM *Plus* standardisation yielding the highest values. But, in general, the results give remarkable confirmation of the quality of the data collected with different tests from different samples and of the accuracy of the conversion tables.

Conclusion

Most of the data presented here suggest that the *Raven's Progressive Matrices* form an excellent series of tests embodying one of the most ingenious ideas test constructors have ever had.

Nevertheless, when interpreting test results, it is important to bear the theoretical basis, purposes, and limitations of the tests in mind. On the one hand, one must ensure that the tests and norms employed are appropriate not only to the group and individuals one wishes to test but also to the purposes for which the testing is being carried out. The results of individual testing must be interpreted in the context of wider information on the person being tested, including information on the tested person's previous experience with tests. Excessive generalisation from the results must be avoided. Deviations from standard procedures must be taken into account etc. (International Test Commission, 2000). On the other hand, failure to fully understand the theoretical basis of the tests and the measurement model employed in their development, has led to a great deal of misguided research and widespread misinterpretation of research results. It will only be if these things are done that the RPM will retain its value, and its unique place in psychology, into the future.

Figure 10.2. The SPM, SPM-*Plus* and APM II converted to SPM (example for age 13)



References

Boben, D. (2003). *Priročnik za Ravnove progresivne matrice in besedne lestvice. Slovenska standardizacija Ravnovih progresivnih matric*: *norme CPM, SPM, APM. [Manual for Raven's Progressive Matrices and Vocabulary Scales. Slovenian Standardization of Raven's Progressive Matrices*: *CPM, SPM, AMP norms.]* Ljubljana: Center za psihodiagnosticna sredstva.

Boben, D. (2005). *Priročnik za Ravnove progresivne matrice in besedne lestvice. SPM-Plus, Slovenske norme za mladostnike v primerjavi z drugimi normami : dodatek 3. zvezka priročnika (SPM). [Manual for Raven's Progressive Matrices and Vocabulary Scales. SPM-Plus*; *Slovenian norms for adolescents in comparison to other norms. Supplement to Section 3.]* Ljubljana: Center za psihodiagnosticna sredstva.

De Reggi, M. (2007). *Psihometrične značilnosti Ravnovih standardnih progresivnih matric, oblika plus, pri slovenskih mladostnikih. Diplomsko delo. [*Psychometric characteristics of Raven's SPM-*Plus* regarding Slovene adolescents. Diploma.*]*Ljubljana: Univerza v Ljubljani. Filozofska fakulteta. Oddelek za psihologijo.

Dobrean (Domuta), A., Comsa, M., Rusu, C. (2005). In R. Balazsi, *Romanian standardization of Raven's Standard Progressive Matrices Plus.* Budimpešta: 8EECPA.

International Test Commission (2000). *International Guidelines for Test Use.*

Jerman, J. (2000). *Ugotavljanje razvoja fonološkega zavedanja pri predšolskih otrocih. Doktorska disertacija. [Development of phonological conscience at pre-school children. PhD Thesis.]* Ljubljana: Univerza v Ljubljani, Pedagoška fakulteta, Oddelek za defektologijo.

Klopčič, A. (2007). *Standardizacija Testa nizov in analiza Flynnovega učinka. Diplomsko delo. [Test Series Standardization and Flynn Effect Analyses. Diploma.]* Ljubljana: Univerza v Ljubljani. Filozofska fakulteta. Oddelek za psihologijo.

Lapajne, Z. (1997). Vlado Schmidt (1910-1997) in prvi slovenski skupinski testi inteligentnosti. [Vlado Schmidt (1910-1997) and first Slovene group intelligence test.] *TESTinfo, novice Centra za psihodiagnosticna sredstva, letnik 2, št. 1*.

Matešić, K. (2000). Relations between results on *Raven Progressive Matrices Plus* sets and school achievement. *Review of Psychology, 7(1-2)*, 75-82.

MacKintosh, N. J. (1998). *IQ and Human Intelligence.* Oxford: Oxford University Press.

Pečjak, V. (1983). *Nastajanje psihologije.* [*Beginning of Psychology.*] Ljubljana: Dopisna delavska univerza Univerzum.

Plut, M. (2003). *Merske lastnosti poskusne verzije Besedne lestvice inteligentnosti Mill Hill. Diplomsko delo. [Psychometric characteristics of experimental version of Mill Hill Vocabulary Scales. Diploma.]* Ljubljana: Univerza v Ljubljani, Filozofska fakulteta, Oddelek za psihologijo.

Raven, J. (1981). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No. 1*: *The 1979 British Standardisation of the Standard Progressive Matrices and Mill Hill Vocabulary Scales, Together With Comparative Date from Earlier Studies in the UK, US, Canada, Germany and Ireland*. San Antonio, TX: Harcourt Assessment.

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1-48.

Raven, J., Raven, J. C. & Court, J. H. (1999a). *Priročnik za Ravnove progresivne matrice in besedne lestvice. 1. zvezek*: *Splošni pregled. [Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1*: *General Overview]* Ljubljana: Center za psihodiagnosticna sredstva.

Raven, J., Raven, J. C., & Court, J. H. (1999b). *Priročnik za Ravnove progresivne matrice in besedne lestvice. 2. zvezek*: *Barvne progresivne matrice. . [Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 2*: *CPM]* Ljubljana: Center za psihodiagnosticna sredstva.

Raven, J., Raven, J. C., & Court, J. H. (1999c). *Priročnik za Ravnove progresivne matrice in besedne lestvice. 3. zvezek*: *Standardne progresivne matrice. . [Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3*: *SPM]* Ljubljana: Center za psihodiagnosticna sredstva.

Raven, J., Raven, J. C., & Court, J. H. (1999d). *Priročnik za Ravnove progresivne matrice in besedne lestvice. 4. zvezek*: *Zahtevne progresivne matrice. [Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 4*: *APM]* Ljubljana: Center za psihodiagnosticna sredstva.

Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3*: *The Standard Progressive Matrices, Iincluding the Parallel and Plus Versions*. San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales, Section 3*: *Standard Progressive Matrices.* San Antonio, TX: Harcourt Assessment.

Rezultati raziskovanj. Seznam šol s splošnimi podatki ob začetku šolskega leta 1995/96. [Research results. List of schools in the beginning of school year 1995/96.] (1996). Ljubljana: Statistični urad Republike Slovenije.

Oakland, T. (1995). 44 country survey shows intenational test use patterns. *Psychology International, 6(1),* Winter, 7.

Statistične informacije [Rapid Reports] (1992). Ljubljana: Statistični urad Republike Slovenije.

Statistične informacije [Rapid Reports] (1997). Ljubljana: Statistični urad Republike Slovenije.

Statistica for Windows. (1995). StatSoft Inc.

Žalik, E. (2003). *Slovenska priredba besedne lestvice izbirnega tipa Mill Hill. Diplomsko delo. [Slovene adaptation of multiple-choice form of the Mill Hill Vocabulary Scales. Diploma.]* Ljubljana: Univerza v Ljubljani, Filozofska fakulteta, Oddelek za psihologijo.

Chapter 11

The Lithuanian Standardisation of the Coloured Progressive Matrices in an International Context*

Gražina Gintilienė, Dovilė Butkienė, and John Raven

Abstract

The *Coloured Progressive Matrices* (CPM) test, developed by J. C. Raven in 1947, is used world-wide to evaluate the non-verbal reasoning ability of 5 to 11 year-old children. This chapter presents Lithuanian CPM norms based on the standardisation conducted in 2004 of a representative sample of 6 to 11 year-old children (n=1067) in an international context: These norms differ, for example, from British and U.S. norms. An Item-Response-Theory-based item analysis, as well as the more usual split-half and internal consistency methods were used to assess the scaleability and validity of the test. It was found that environmental factors (place of residence of child and educational level of the parents) had a more significant effect in the development of non-verbal reasoning ability than the gender of the child. It is possible to conclude that both the test itself and the norms that were developed will be useful in evaluating the non-verbal reasoning ability of children for screening purposes.

*****

Raven's *Progressive Matrices* (RPM) are a series of measures which assess the ability to find meaning in confusing situations although they are often said to measure non-verbal reasoning ability. Already in 1930, while conducting research with mentally retarded subjects, J.C.

* The authors wish to thank Mrs Jean Raven for her help in generating of the smoothed percentiles.
We also thank Vida Gudauskiene, the psychologist of National Centre of Special Education and Psychology for her assistance in organisation of test administration at schools. Also we are grateful to psychologists of psychological pedagogical services for their voluntary work testing children and also to all Vilnius University masters degree educational psychology students participated in this study during the academic year 2004-2005.

Raven understood the need to develop a method which would allow the investigation not only of genetic factors affecting intelligence, but also environmental ones. That motivated him to develop a test whose results would be minimally dependent on acquired information but would be theoretically based and unambiguously interpretable. The initial version of the RPM appeared in 1938. Its theoretical base was Spearman's "***g***" factor which consists of two components: eductive and reproductive abilities. The term "eductive" comes from the Latin verb "educere" and means "the ability to make meaning out of confusion". The second term encompasses the ability to reproduce acquired information (Raven, 2000). Eductive ability includes the perception of a problematic situation and its analysis, the detection of problem, going beyond the given to perceive that which is not immediately obvious, forming constructs which make it easier to think about complex problems involving many mutually dependent variables. In the other words, it is the process of making sense of novel complex situations, when conclusions are based on active purposeful discovery and new insights rather than on simple choice among presented observable options. In solving the RPM the subject has to identify the missing detail of the picture, after determining the relationship among the elements presented in the matrix. The tasks become progressively more difficult. Reproductive ability encompasses the skill of learning, recalling, and reproducing verbal material. It is evaluated using Vocabulary scales, which are part of Raven's test battery (Raven, J., Raven, J.C., Court, 1998a).

The first version of RPM to be published was the *Standard Progressive Matrices* (SPM). Although the *Coloured Progressive Matrices* was developed at the same time, it was not published until 1947, at which time the *Advanced Progressive Matrices* (APM) was also published (although it had been developed for use by the armed services during the Second World War). The CPM comprises Sets A and B of the SPM with an additional set – Set Ab – interpolated between them. All CPM items are in colour. At this time, there are four versions of the RPM in use: CPM, SPM, APM and SPM*Plus* (SPM+). The SPM+ was developed to restore the discriminative power which the SPM had when it was first developed but which had been eroded by the so-called "Flynn effect". In 1998 parallel versions of SPM and CPM were developed to foil respondents who might have been coached in the correct answers (Raven et al., 1998a).

In developing the RPM, J. C. Raven tried to minimize the dependence of task solution on acquired knowledge, especially that acquired through

formal education. In research he carried out in 1936 he showed that the test worked – scaled – in the same way for children from all backgrounds. Nevertheless, there were major differences between the means and spread of the actual scores obtained by children from different socio-economic and educational backgrounds. A comparison of the 1979 and 1938 normative data for the *Standard Progressive Matrices* revealed that there had been a marked increase in the scores obtained by young people of the same age (Raven, 1981). Later, in 1987, Flynn reported that the population IQs of people from many countries had increased by 5 to 25 points from generation to generation. Moreover, the difference in mean IQ between generations was greater when non-verbal tests (such as the RPM) were used rather than when the assessments were made using multi-component tests such as the Wechsler or Stanford-Binet tests. That motivated researchers to look for other factors possibly influencing the results of the RPM. The latest research (at present the RPM have been standardised in more than 30 countries) does not provide a simple answer regarding factors influencing the development of reasoning abilities. In the opinion of Raven, J. (2000b) growth on the RPM score mean may be linked to the same factors as parallel increases in human height, birth weight, life expectancy, athletic ability, and the drop in mortality of infants - that is, to better nutrition, welfare, and hygiene. In addition, he noted that certain child-rearing practices and forms of education have a significant effect on the development of children's eductive ability. Those factors emphasise the importance not only of having separate standardisations for various countries, but also the need to re-evaluate the norms periodically because a country's intellectual potential changes with its economic and technical development. In fact, Aiken (2003) states that restandardisation is needed for all three RPM, including the CPM.

Until now not a single RPM version was standardised in Lithuania on a representative sample of the population. In 1999 259 children were evaluated in Kaunas, the second biggest town in Lithuania, using the CPM (Lynn & Kazlauskaite, 2002). Despite the fact that the study was not representative, being based on the results of children from some city schools only, it was concluded that the IQ of Lithuanian children is 94 and is lower than that of Russian children (IQ=97) and Estonian children (IQ=99). However, the first intelligence test, Wechsler Intelligence Scale for Children - third edition (WISC-III) was standardised in Lithuania only in 2002.

This unsatisfactory lack of standardised instruments prompted the current undertaking, which was intended to establish Lithuanian normative

data for, and examine the psychometric properties of, Raven's *Coloured Progressive Matrices* among 6 to 11 year-olds.

Methods

Participants

To select a representative sample of 6-11 year old Lithuanian children, several random sampling methods were used. First, by means of a stratified sampling method, centred specifically on such variables as place of residence (urban, town, or rural), native language (Lithuanian, Russian, or Polish), and the kind of educational setting (kindergarten or mainstream school), 79 educational institutions were selected. Further, out of each selected educational institution, by means of a simple random sampling method, 1-2 groups, or primary-school first classes (2nd, 3rd, and 4th ones) were selected. Finally, out of each selected class (group), based on its size, by means of a random sampling method, 2-6 children (boys and girls in equal proportion) fulfilling the age requirements for representative sample were selected. Demographic characteristics of representative sample (n=1067) are presented in Table 11.1. The distribution of children in total sample (also in each age group) by place of residence and gender closely corresponds with the data provided by the Lithuanian Statistics (Education, 2004).

The sample consists of 11 age groups, each of them covering one half-year. For example, the 6-year old group includes children whose age ranges from 5 years 9 months and 1 day to 6 years 2 months and 30 days. Similarly, the 6 ½ year olds' group includes children whose age ranges from 6 years 10 months and 1 day to 6 years 8 months and 30 days. Children of the same age, quite often, attend different grades, which is also true in case of 6- or 7-year olds among whom one may find preschoolers attending a kindergarten, and pupils attending 1st or 2nd grade at a primary school. Thus, the selection of children among selected class (group) pupils was based on the following criterion: The child had to be not younger than 5 years 9 months yet not older than 11 years 2 months. Table 11.2 presents the data on educational institutions attended by representative sample children, and on grades to which they belong.

Additional information on children's native language, educational programme, and parents' education was gathered from children's parents (caregivers) by means of a questionnaire. According to parents (caregivers) report, 87.7% children spoke only Lithuanian, 5.2% only

Table 11.1. *Coloured Progressive Matrices*
Demographic Characteristics of Lithuanian Sample

		Place of Residence						Gender			
		Urban		Town		Rural		Boys		Girls	
	n	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Age group											
6	50	21	42.0	16	32.0	13	26.0	25	50.0	25	50.0
6½	93	38	40.9	29	31.2	26	28.0	46	49.5	47	50.5
7	88	36	40.9	26	29.5	26	29.5	44	50.0	44	50.0
7½	98	40	40.8	29	29.6	29	29.6	49	50.0	49	50.0
8	88	36	40.9	26	29.5	26	29.5	44	50.0	44	50.0
8½	122	49	40.2	37	30.3	36	29.5	61	50.0	61	50.0
9	122	50	41.0	36	29.5	36	29.5	61	50.0	61	50.0
9½	115	47	40.9	34	29.6	34	29.6	58	50.4	57	49.6
10	95	39	41.1	28	29.5	28	29.5	48	50.5	47	49.5
10½	111	45	40.5	33	29.7	33	29.7	56	50.5	55	49.5
11	85	35	41.2	25	29.4	25	29.4	42	49.4	43	50.6
Gender											
Boys		218	40.8	160	29.9	156	29.3				
Girls		218	40.9	159	29.8	156	29.3				
Total	1067	436	40.9	319	29.9	312	29.2	534	50.0	533	50.0
Lithuanian population*			41.7		32.5		25.8				
Lithuanian population**			40.7		29.8		29.5				

* Percentage of 6-7 years old children attending preschool education groups at mainstream schools or kindergartens according their residential areas (Education, 2004).

** Percentage of children attending grades 1-4 in urban, town, and rural areas. (*Education*, 2004).

Russian, 3.7% only Polish, and 3.3% – several languages at home. Such distribution of children under study by native language corresponds with the data provided by the Lithuanian Statistics, 2004, on the number of Lithuanian children receiving education in different languages of instruction (Education, 2004).

Information on parents' education was reported by 934 (87.5% of the total number of children under study) parents (caregivers). Almost half the parents indicated they had post-secondary or higher education (44.6% fathers and 51.2% mothers). 5-6% of the parents' education was

Table 11.2. *Coloured Progressive Matrices*
Lithuanian Sample by Educational Setting

		Not attending kindergarten / school		Kindergarten		School							
						Grade 1		Grade 2		Grade 3		Grade 4	
	n	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
Age group													
6	50	4	8.0	45	90.0	1	2.0						
6½	93	6	6.5	78	83.9	9	9.7						
7	88	5	5.7	34	38.6	48	54.5	1	1.1				
7½	98	2	2.0			86	87.8	10	10.2				
8	88					56	63.6	32	36.4				
8½	122					11	9.0	100	82.0	11	9.0		
9	122					2	1.6	80	65.6	40	32.8		
9½	115							10	8.7	101	87.8	4	3.5
10	95							2	2.1	71	74.7	22	23.2
10½	111									10	9.0	101	91.0
11	85					2	2.4			3	3.5	80	94.1
Total	1067	17	1.6	157	14.7	215	20.1	235	22.0	236	22.1	207	19.4

basic. The rest indicated that they graduated secondary or vocational school.

A representative sample of 6-11 year olds was selected including special needs children in mainstream schools. According to data provided by parents (caregivers), the majority of pupils attending 1st-4th grades were following a general education curriculum (96.1%), however the sample included pupils who studied according modified (2.4%), adapted (1.3%), or special individual (0.1%) programmes.

Description of the Coloured Progressive Matrices

The CPM consists of 36 coloured and attractive items: Three Sets (A, Ab, and B), each of 12 items, representing a series of patterns with a bit missing. Sets A and B correspond to the SPM Sets A and B, with Set Ab – drawn up for the CPM version – between them. Success in Set A is determined by individual's ability to complete a continuous pattern changing, first, in one direction, then, at the end of the set, in two directions at the same time. Success in Set Ab depends on individual's

ability to conceive discrete figures as a spatially related whole, and to choose a figure which completes the pattern. Set B is composed of items demanding ability to reason by analogy. Each of the 36 items requires that the child find a missing detail among six alternatives given at the foot of the picture, and to indicate it. The CPM is designed to assess the eductive ability of 5-11 year olds, mentally retarded people, and the elderly. By "eductive ability" is meant the ability to make meaning out of confusion; the ability to perceive. This involves simultaneously forging images of wholes and parts (this is not correctly rendered as "perceiving relationships between elements", because elements do not exist without wholes). It also involves the ability to reason by analogy. Some years after the first publication of the CPM, additional item analyses led to the modification of certain items and the order of presentation (Raven et al., 1998a).

Data on the reliability and validity of the CPM are available in the test Manual (Raven et al., 1998a). Split-half reliability estimates range from 0.65 to 0.94 (age groups from 6 to 8 years); retest reliabilities from 0.71 to 0.87 (5.7 and 8 year groups); and Cronbach alphas from 0.80 to 0.93 (5 and 11 ½ year age groups).

The validity of the CPM is supported by statistically significant correlations (from 0.50 to 0.80) with other "intelligence" tests (Terman-Merill, WISCR-R, Stanford-Binet) and achievement tests (Raven et al., 1998a), by scaleogram analyses and factor analyses of the correlations between numerous tests (e.g. Snow, Kyllonen and Marshalek, 1984, and Carroll, 1993) but, most importantly, by internal consistency analyses carried out on the test itself using Item Response Theory. As can be seen from other chapters in this volume, these show that the RPM tests have many of the properties of a "tape measure" or "meter stick" and that these are almost constant across many cultural groups.

Procedure

The standardisation was carried out in February – March and November – December, 2004. With the approval of Lithuanian Republic Ministry of Education and Science, the administration of educational institutions was informed about the study. Teachers were asked to hand a letter and questionnaire to children's parents (caregivers). The questionnaire was designed to obtain socio-demographic data (parents' education, child's native language, family composition), and other supplementary information (attending or non-attending a pre-school institution before

entering school proper, educational programme, additional education, etc.). Each child was tested with the prior written permission issued by his/her parents (caregivers).

Thirty eight specially trained primary researchers administered the booklet form of the CPM to selected children. The testing was conducted during morning hours at school in a room provided for the testing purpose by the manager of the educational institution. Kindergarteners and pupils attending 1st grade (like the children who did not attend any educational setting) were tested individually. Pupils attending 2nd grade were tested in pairs. Pupils attending 3rd and 4th grades were tested in groups including 4-6 children. The CPM administration took about 10 to 20 minutes.

As indicated in CPM Manual (Raven et al., 1998a), children unable to succeed on the first five items of Set A fail to realise how the test items are to be solved. As a result, their final score on CPM, whatever it may be, is considered to be invalid. So, the further data analysis was based only on results of those children who accomplished the first five tasks on the CPM.

Results

Norms

Means and standard deviations were generated at half a year intervals between ages 5 years and 9 months to 11 years 2 months. In order to eliminate sampling errors the raw scores for the main percentiles (5^{th}, 10^{th} etc.) for each age group were plotted and then smoothed by graphing across age groups. The Table 11.3 presents smoothed CPM norms for Lithuania.

There is a clear ceiling effect among older children (9 ½ +), and, even at the younger ages, the distributions are not symmetrical around the median score. For example, among 7 ½ year olds, the difference between the 5^{th} and 50^{th} percentile is 9 while that between the 50^{th} and 95^{th} is only 6. Because the test has only 36 items, the scores of the most able 10% of the population seem not to increase from 9 ½ years of age to 11 years of age.

It is obvious from these results that the test is unsuitable for use with more able children over nine years of age. As indicated in the Manual, those who wish to obtain reliable information on such children should (once their superior performance on the CPM has become apparent)

Table 11.3. *Coloured Progressive Matrices*
Smoothed Percentile Norms for Lithuania for 2004

	Age in Years (Months)										
	6	**6½**	**7**	**7½**	**8**	**8½**	**9**	**9½**	**10**	**10½**	**11**
	5(9)	6(3)	6(9)	7(3)	7(9)	8(3)	8(9)	9(3)	9(9)	10(3)	10(9)
	to	to	to	to	To	to	to	to	to	to	to
Percentile	6(2)	6(8)	7(2)	7(8)	8(2)	8(8)	9(2)	9(8)	10(2)	10(8)	11(2)
95	27	29	30	31	32	33	34	35	35	35	35
90	25	27	29	30	31	32	33	34	34	34	34
75	23	24	26	27	28	30	31	32	33	33	33
50	**21**	**22**	**23**	**25**	**26**	**27**	**28**	**29**	**30**	**30**	**30**
25	17	18	19	20	21	22	24	25	26	27	27
10	14	15	16	17	18	19	20	21	22	23	24
5	13	14	15	16	17	18	19	20	21	22	22
n	*50*	93	88	98	88	122	122	115	95	111	85

* Generated in collaboration with Jean Raven.

proceed to the *Standard* Progressive Matrices on completion of Set B and deduct the scores for Set Ab from the final scores obtained before seeking to convert them to percentile scores.

Group Differences

Table 11.4 compares the overall mean scores of boys and girls, assessing the statistical significance of the differences using the t test for independent samples. It is clear that there is no difference in the overall scores this group of young people aged 6 to 11 years.

An analysis of variance revealed significant differences by place of residence ($F(3, 1064) = 26{,}09$, $p < 0.001$, for raw scores and $F(3, 1064) = 36.26$, $p < 0.001$ for percentiles). As shown in Table 11.5 there are significant differences between children from urban areas and town or rural groups, with the children from the cities performing better than their peers attending town and village schools. The mean difference between last two groups is not statistically significant, although children from the towns do slightly better than rural children.

The analysis of the variation in scores with parental education was based on parents' answers to the questionnaire they had completed

Table 11.4. *Coloured Progressive Matrices*
Lithuanian Standardisation
Means and Standard Deviations by Gender

	Mean	Standard deviation	t	p
Raw scores				
Boys	26.34	5.63	1.151	0.250
Girls	25.94	5.64		
Percentile				
Boys	51.90	28.60	1.102	0.271
Girls	49.92	29.85		

Table 11.5. *Coloured Progressive Matrices*
Lithuanian Standardisation
Means and Standard Deviations by Place of Residence

	Mean	Standard deviation	t	p
Raw scores				
Urban	27.55	5.04	4.935	0.000
Town	25.57	5.93		
Urban	27.55	5.04	7.133	0.000
Rural	24.74	5.68		
Town	25.57	5.93	1.809	0.071
Rural	24.74	5.68		
Percentile				
Urban	59.38	26.96	5.479	0.000
Town	47.91	29.44		
Urban	59.38	26.96	8.257	0.000
Rural	42.15	28.97		
Town	47.91	29.44	2.478	0.013
Rural	42.15	28.97		

Table 11.6. *Coloured Progressive Matrices*
Lithuanian Standardisation
Means and Standard Deviations by Educational Level of Parents

Type of school (years of school completed)	Father					Mother				
		Raw Scores		Percentile			Raw Scores		Percentile	
	n	M	SD	M	SD	*n*	M	SD	M	SD
Basic (10 years)	46	22.98	5.67	33.70	25.44	57	22.68	5.82	31.98	25.82
Secondary (12 years)	170	24.86	5.83	42.74	28.41	211	24.82	5.60	42.21	28.11
Vocational (11-13 years)	265	24.88	5.53	45.09	29.12	180	24.85	5.64	46.93	29.45
Post-secondary (14-15 years)	210	26.90	5.28	55.34	27.06	257	26.30	5.55	51.67	28.81
Higher (college, university) (16-18 years)	177	28.07	5.25	66.27	26.62	224	28.04	5.17	63.99	26.52

Table 11.7. *Coloured Progressive Matrices*
Lithuanian Standardisation
Significance of Differences Between Mean Scores of Groups by Different Educational Level of Parents (see Table 11.6)

	Father's Educational level									
	B / S	B / V	B/ P	B / H	S / V	S / P	S / H	V / P	V / H	P / HI
Raw score										
t value	1.959	2.141	4.509	5.761	0.019	3.576	5.384	4.054	6.070	2.165
p value	0.051	0.033	0.000	0.000	0.985	0.000	0.000	0.000	0.000	0.031
Percentile										
t value	1.956	2.493	4.971	7.460	0.831	4.424	7.966	3.940	7.752	3.979
p value	0.052	0.013	0.000	0.000	0.406	0.000	0.000	0.000	0.000	0.000

	Mother's Educational level									
	B / S	B / V	B / P	B / H	S / V	S / P	S / H	V / P	V / H	P / H
Raw score										
t value	2.534	2.508	4.416	6.800	0.053	2.866	6.228	2.677	5.912	3.525
p value	0.012	0.013	0.000	0.000	0.958	0.004	0.000	0.008	0.000	0.000
Percentile										
t value	2.480	3.437	4.752	8.179	1.619	3.572	8.315	1.676	6.116	4.854
p value	0.014	0.001	0.000	0.000	0.106	0.000	0.000	0.094	0.000	0.000

B – Basic, S – Secondary, V- Vocational, P – Post-Secondary, H – Higher

about the school programme they had completed. ANOVA revealed significant effects for fathers ($F(4, 864) = 15.79$, $p < 0{,}001$, for raw scores and $F(4,864) = 25.84$, $p < 0.001$ for percentiles), and mothers ($F(4, 929) = 17.29$, $p < 0.001$, for raw scores and $F(4,929) = 24.58$, $p < 0.001$ for percentiles) education. Data presented in the Tables 11.6 and 11.7 show that the mean raw score of the pupils increases with parents' educational level. The differences between the groups are statistically significant. ($p<0.05$) except for groups of children whose fathers and mothers completed vocational and secondary schools.

Item Analyses

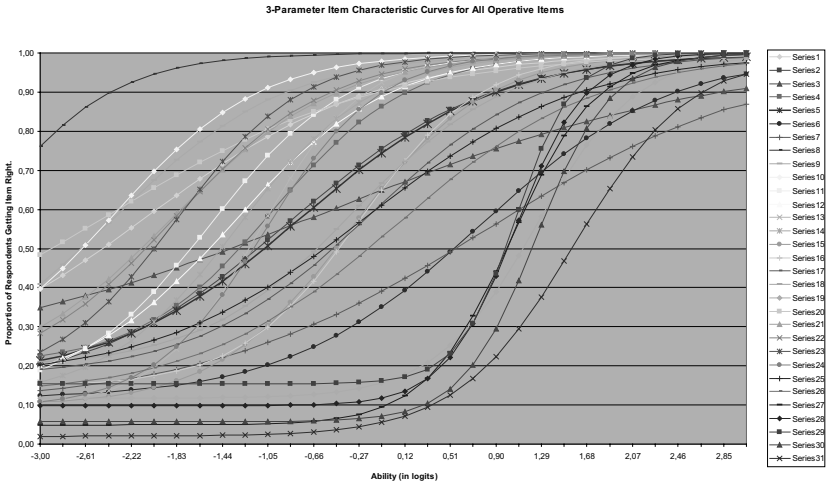
Split-half internal consistency correlations were calculated for the total sample and each separate age group and are presented in Table 11.8. All were high and ranged from 0.82 (seven year olds) to 0.90 (eight and half year olds and total sample). The internal consistency was also assessed using Cronbach's alpha (see Table 11.8). It can be seen that the Cronbach alpha coefficients are comparable to the split-half estimates: the lowest (0.76) was for six and seven year olds and the highest (0.84) was for the total sample.

As explained in the *General* Section of the RPM *Manual* (Raven et al., 1998, updated 2003) and in the *General Introduction* to this book, it does not make a great deal of sense to intercorrelate the items of tests developed according to Item Response Theory (IRT) and then use those correlations to calculate statistics like Cronbach's Alpha, still less as a basis for factor analyses in which an attempt is made to assess the "unidimensionality" of a test: Consider how meaningful it would be to conduct the same exercise based on the centimetre marks on a tape measure.

Table 11.8. *Coloured Progressive Matrices*
Lithuanian Standardisation
Split Half and Cronbach Alpha Coefficients for Total Sample and by Age Group

		Age Group										
	All	6	6½	7	7½	8	8½	9	9½	10	10½	11
Split Half	0.90	0.82	0.86	0.81	0.84	0.86	0.90	0.87	0.89	0.84	0.82	0.84
α	0.87	0.76	0.81	0.76	0.80	0.82	0.86	0.86	0.85	0.82	0.80	0.80

Figure 11.1. *Coloured Progressive Matrices*
Lithuanian Standardisation
3-Parameter Item Characteristic Curves for All Operative Items



Joerg Prieler was therefore commissioned to undertake a graphical 3-parameter Item-Response-Theory-based analysis along the lines discussed in other chapters of this book.

Figure 11.1 displays the Item Characteristic Curves for all the 31 items of the test that remained after the first five had been discarded because, as explained earlier, all respondents who got any of these items wrong were rejected from the analysis on the grounds that they had not understood what they were supposed to do.

It is immediately obvious that one item ... Number 3 in this plot, but which is actually A8 ... is seriously defective: Far too many low ability pupils get it right and far too many able pupils get it wrong. (It would not, however, be appropriate to modify this item in the test at this point in time since doing so would invalidate future comparisons with the vast amount of international and historical data that have been accumulated with the test as it stands.)

Also striking is the cluster of four items (three of which are Items 9, 10, and 11 from Set B) whose ICCs rise steeply toward the right hand side of the plot. What these tell us is that less able pupils hardly ever get these items right (except by "chance"), but, once pupils have developed the abilities needed to understand them, they seldom get them wrong.

Figure 11.2. *Coloured Progressive Matrices*
Lithuanian Standardisation
Test Characteristic Curve Calculated Over All Operative Items

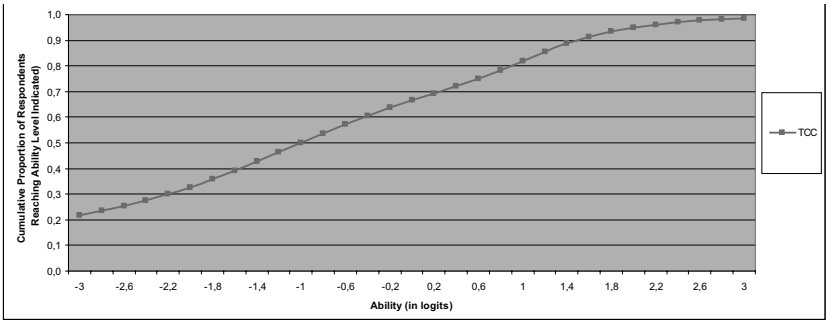
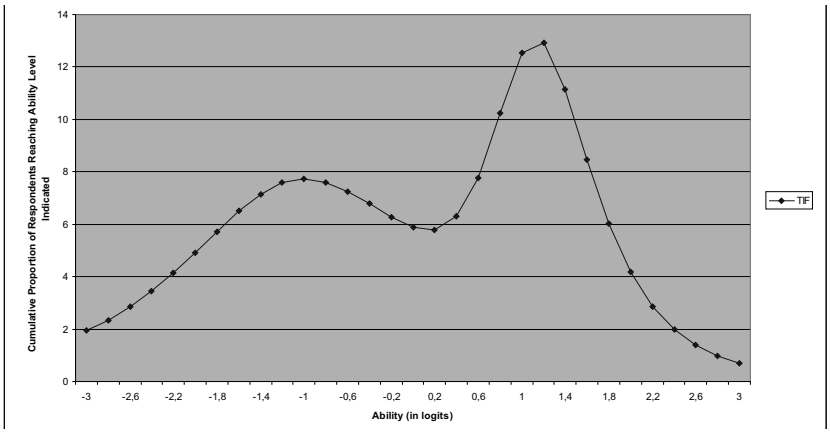


Figure 11.3. *Coloured Progressive Matrices*
Lithuanian Standardisation
Test Information Curve Over Operational Range of Test



It is also notable that that there are few items discriminating well in the middle ability range. The implications of this will become clear in a moment.

Figure 11.2 shows the Test Characteristic Curve for the Test as a whole. Whereas most people would expect it (and think it "ought") to approximate a Gaussian (often misleadingly termed "normal") ogive, it is surprisingly linear with few leaps or plateaux.

Figure 11.3 shows the Test Information Function curve.

At this point, a brief, if only approximate, explanation of what the Test Information Function (TIF) curve tells us may be offered (a suc-

cinct, but comprehensible and powerful, explanation will be found in Hambleton, et al., 1991). Basically, if there are many items with good discriminatory power (i.e. steep ICCs) at a particular level of ability covered by the test, the test yields a great deal of good information enabling users to accurately discriminate between those having that level of ability. If there are few items with good discrimination indices (i.e. having steep ICCs) at a particular point … usually at the top and bottom end of the range of ability for which the test is intended … little reliable diagnostic information can be gained from using the test. Thus if, as is often the case, users require a test which discriminates among those who appear to have particularly high or low ability, a test having the typical Gaussian-shaped Test Information Function curve will not give them what they want. A test having a rectilinear, or even bimodal, TIF would be of greater value. (In point of fact, of course, since the overall TIF does not reflect what happens within age groups, it would be desirable to calculate them separately within age groups. As it happens, however, as indicated above, a close inspection of percentile norms presented within age groups enables one to derive the relevant information directly.)

Be that as it may, the shape of the Test Information Function curve for the CPM shown above will disconcert many people. But, far from suggesting that the authors should tinker with the test to obtain a more Gaussian curve, the preceding discussion suggests that, if an attempt were to be made to improve the test, attention should concentrate on getting better discrimination in the upper and lower ability ranges.

Validity

Given normal cognitive development in children, one indication of the validity of the CPM would be a linear progression in raw scores over the age groups tested. The relationship between age and mean CPM raw scores was assessed by calculating a Pearson product-moment correlation coefficient. This was 0.52, thus accounting for 27.4% of the overall CPM total score distribution. As it is shown in Table 11.9 the mean of raw scores of 6-11 year old children increase with age (with the exception of ten and half and eleven year olds groups where, because of the ceiling effect noted above, more able children were unable to demonstrate their abilities).

Before the IRT based ICC analysis reported above had been commissioned, we assessed the content validity of the test on the basis of conventional item difficulty indices. Despite the more detailed information available from the plots of the ICCs, our analysis in terms of item difficulty

Table 11.9. *Coloured Progressive Matrices*
Lithuanian Standardisation
Means and Standard Deviations by Age Group

Total	Age Group										
	6	6½	7	7½	8	8½	9	9½	10	10½	11
Total											
M	19.94	21.29	23.23	23.61	24.78	26.08	27.77	28.09	28.85	30.01	29.41
SD	4.30	4.75	4.22	4.82	4.97	5.50	5.21	5.10	4.50	4.15	4.30
Set A											
M	8.34	8.88	9.22	9.32	9.44	9.61	10.04	10.07	10.36	10.46	10.38
SD	1.47	1.39	1.39	1.42	1.26	1.46	1.56	1.47	1.14	1.11	1.25
Set Ab											
M	6.58	7.28	8.11	8.09	8.58	9.23	9.63	9.75	10.08	10.46	10.34
SD	2.20	2.25	1.94	2.24	2.39	2.43	2.13	2.48	1.91	1.71	1.56
Set B											
M	5.02	5.13	5.90	6.20	6.76	7.25	8.10	8.27	8.41	9.09	8.69
SD	1.81	1.92	1.89	2.15	2.47	2.66	2.52	2.44	2.48	2.26	2.30
n	50	93	88	98	88	122	122	115	95	111	85

indices may nevertheless be of interest. The data presented in Table 11.10 reveal that there are some exceptions to a smooth progression in item difficulty. Items A9, Ab7, and B10 were slightly easier than one would expect and were more successfully solved than earlier items in correspondent Sets by children more than eight years old.

Discussion

In this paper we have reported the results of the first standardisation of the CPM on a Lithuanian representative sample. Internal consistency assessed by split half and Alpha coefficients are similar to those obtained elsewhere (Raven et al., 1998a).

However, as can be seen from Table 11.11, the Lithuanian norms do, however, differ from what has, in effect, become the international standard – the 1982 British norms - (Raven et al., 1998a). (They differ even more from the 1986 U.S. norms, which are low by international standards [Raven, 2000a]). If Lithuanian children were evaluated using the UK or US norms then, for example, the average 6 ½ year old would appear to score as well as the average British or American 8 year old. This does not necessarily show that Lithuanian children are more able than their British and American counterparts, but is probably due to the

Table 11.10. *Coloured Progressive Matrices*
Lithuanian Standardisation
Item Difficulties (p values) by Age Group

	Age Group											
Item	6-11	6	6½	7	7½	8	8½	9	9½	10	10½	11
A1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
A2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
A3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
A4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
A5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
A6	0.96	0.94	0.94	0.97	0.96	0.94	0.93	0.94	0.97	1.00	1.00	0.99
A7	0.79	0.62	0.63	0.73	0.70	0.85	0.75	0.86	0.81	0.86	0.90	0.88
A8	0.76	0.56	0.69	0.78	0.74	0.68	0.71	0.75	0.77	0.90	0.84	0.81
A9	0.80	0.42	0.63	0.69	0.70	0.71	0.85	0.88	0.91	0.91	0.95	0.91
A10	0.77	0.54	0.65	0.72	0.68	0.68	0.74	0.82	0.86	0.86	0.88	0.89
A11	0.35	0.14	0.15	0.16	0.30	0.28	0.33	0.46	0.41	0.47	0.51	0.48
A12	0.30	0.12	0.19	0.17	0.24	0.30	0.30	0.34	0.35	0.36	0.39	0.41
Ab1	0.99	0.92	1.00	1.00	1.00	0.99	0.98	1.00	0.98	1.00	1.00	1.00
Ab2	0.97	0.94	0.96	0.96	0.95	0.98	0.97	0.98	0.97	0.95	0.97	1.00
Ab3	0.96	0.94	0.95	0.99	0.93	0.99	0.93	0.97	0.94	0.95	1.00	0.98
Ab4	0.87	0.74	0.71	0.86	0.79	0.82	0.84	0.92	0.90	0.95	0.96	0.94
Ab5	0.85	0.68	0.73	0.77	0.87	0.88	0.85	0.88	0.87	0.87	0.92	0.92
Ab6	0.76	0.44	0.46	0.60	0.71	0.77	0.78	0.82	0.81	0.88	0.92	0.94
Ab7	0.83	0.50	0.61	0.81	0.75	0.80	0.84	0.90	0.87	0.94	0.97	0.98
Ab8	0.63	0.28	0.33	0.39	0.46	0.55	0.73	0.77	0.72	0.82	0.80	0.75
Ab9	0.62	0.26	0.34	0.50	0.53	0.50	0.64	0.67	0.76	0.75	0.83	0.82
Ab10	0.63	0.38	0.44	0.51	0.45	0.56	0.70	0.68	0.74	0.78	0.78	0.72
Ab11	0.66	0.36	0.56	0.55	0.53	0.56	0.68	0.70	0.76	0.77	0.80	0.82
Ab12	0.32	0.14	0.18	0.18	0.13	0.21	0.29	0.35	0.44	0.43	0.50	0.47
B1	0.98	0.96	1.00	0.97	0.98	0.99	0.98	0.96	0.98	1.00	0.98	0.99
B2	0.93	0.86	0.83	0.91	0.89	0.93	0.92	0.98	0.94	0.97	1.00	0.94
B3	0.92	0.80	0.81	0.86	0.92	0.91	0.92	0.96	0.93	0.99	0.98	0.94
B4	0.90	0.86	0.76	0.85	0.87	0.91	0.89	0.95	0.91	0.94	0.94	0.95
B5	0.76	0.46	0.54	0.66	0.65	0.73	0.76	0.86	0.88	0.83	0.87	0.89
B6	0.64	0.32	0.51	0.56	0.53	0.56	0.61	0.75	0.71	0.73	0.80	0.69
B7	0.55	0.40	0.31	0.43	0.44	0.50	0.55	0.59	0.67	0.67	0.69	0.65
B8	0.33	0.04	0.05	0.09	0.21	0.24	0.30	0.41	0.46	0.47	0.55	0.54
B9	0.38	0.10	0.10	0.13	0.22	0.32	0.36	0.48	0.53	0.55	0.61	0.58
B10	0.46	0.12	0.14	0.26	0.26	0.32	0.47	0.56	0.63	0.58	0.73	0.67
B11	0.32	0.08	0.07	0.10	0.15	0.25	0.33	0.39	0.42	0.44	0.59	0.52
B12	0.17	0.02	0.02	0.08	0.08	0.11	0.16	0.20	0.21	0.24	0.33	0.33

internationally established intergenerational increase in scores (which has become known as the "Flynn effect) that has probably continued from 1982 into the present.

Although there was a tendency for boys to do better on the CPM than girls in the current study, no significant gender differences were found. This supports the British finding that the correlation of 0.01 between the SPM and sex is minimal (Raven et al., 1998b). In this context we may report that no significant differences were found between genders in the Performance IQ of the WISC-III, although the boys' Perceptual Organisation Index was significantly higher (Gintiliene & Girdzijauskiene, 2000). The results of the current investigation confirm previous findings that the same norms can be used in evaluating both boys and girls and that it is enough to maintain an equal number of males and females at various age groups in the standardisation sample.

In choosing the representative sample, a factor which was taken into account was place of residence. When the results of children living in different areas were examined, significant differences were found between those residing in large cities and those in rural areas. Similar results were found with the WISC-III (Gintiliene and Girdzijauskiene, 2003) when samples from rural and urban schools were matched on variables of gender, age and parents educational level. The Full scale IQ of rural children remained significantly lower. These findings may be related to income. The disposable income of rural population is 1.4 times smaller than that of residents living in large cities (Household Income and Expenses, 2004). These differences result in unequal opportunities for parents to develop children's abilities. In addition, the educational level of the parents is another socio-economics factor influencing the results. The WISC-III standardisation results revealed a statistically significant correlation ($r=0.32$, $p<0.01$) between a child's IQ and the educational level of his parents (Gintiliene and Girdzijauskiene, 2003). This relationship also has been pointed out by Sattler (2001) and others. The current research confirmed the fact that the lower the parents' educational level was, the lower were the child's CPM results. No difference between CPM results of children whose parents completed secondary school and those whose parents finished vocational school was found. This may be due to similarity of curricula between them. This fact also confirms CPM results' dependence on parents' educational level.

Table 11.11. *Coloured Progressive Matrices*
2005 Lithuanian Norms in the Context of 1982 British and 1986 American Norms

	Age in Years (Months)											
		6			6½			7			7½	
	5(9) to 6(2)	5(9) to 6(2)	5(9) to 6(2)	6(3) to 6(8)	6(3) to 6(8)	6(3) to 6(8)	6(9) to 7(2)	6(9) to 7(2)	6(9) to 7(2)	7(3) to 7(8)	7(3) to 7(8)	7(3) to 7(8)
Percentile	UK	US	**LT**	UK	US	**LT**	UK	US	**LT**	UK	US	**LT**
95	24	25	**27**	26	28	**29**	28	30	**30**	31	31	**31**
90	21	23	**25**	23	25	**27**	25	27	**29**	28	29	**30**
75	19	19	**23**	20	21	**24**	21	23	**26**	23	25	**27**
50	**16**	**14**	**21**	**17**	**16**	**22**	**18**	**18**	**23**	**20**	**20**	**25**
25	13	12	**17**	14	13	**18**	16	14	**19**	17	15	**20**
10	11	10	**14**	12	11	**15**	13	12	**16**	14	13	**17**
5	9	9	**13**	11	9	**14**	12	10	**15**	13	11	**16**
n	23		**50**	42		**93**	54		**88**	55		**98**

	Age in Years (Months)											
		8			8½			9			9½	
	7(9) to 8(2)	7(9) to 8(2)	7(9) to 8(2)	8(3) to 8(8)	8(3) to 8(8)	8(3) to 8(8)	8(9) to 9(2)	8(9) to 9(2)	8(9) to 9(2)	9(3) to 9(8)	9(3) to 9(8)	9(3) to 9(8)
Percentile	UK	US	**LT**	UK	US	**LT**	UK	US	**LT**	UK	US	**LT**
95	32	32	**32**	33	33	**33**	34	34	**34**	35	35	**35**
90	30	30	**31**	32	31	**32**	33	32	**33**	33	33	**34**
75	25	27	**28**	27	29	**30**	29	30	**31**	31	31	**32**
50	**22**	**22**	**26**	**24**	**24**	**27**	**26**	**26**	**28**	**28**	**27**	**29**
25	18	17	**21**	20	19	**22**	22	21	**24**	24	22	**25**
10	15	14	**18**	16	15	**19**	17	16	**20**	19	17	**21**
5	14	12	**17**	14	12	**18**	15	13	**19**	16	14	**20**
n	44		**88**	48		**122**	52		**122**	37		**115**

(continued)

Table 11.11. *Coloured Progressive Matrices*
2005 Lithuanian Norms in the Context of 1982 British and 1986 American Norms *(continued)*

	Age in Years (Months)								
	10			10½			11		
	9(9) to 10(2)	9(9) to 10(2)	9(9) to 10(2)	10(3) to 10(8)	10(3) to 10(8)	10(3) to 10(8)	10(9) to 11(2)	10(9) to 11(2)	10(9) to 11(2)
Percentile	UK	US	**LT**	UK	US	**LT**	UK	US	**LT**
95	35	35	**35**	35	35	**35**	35	35	**35**
90	33	33	**34**	34	34	**34**	35	34	**34**
75	32	32	**33**	33	32	**33**	33	33	**33**
50	**30**	**28**	**30**	**31**	**29**	**30**	**31**	**30**	**30**
25	25	23	**26**	26	24	**27**	28	25	**27**
10	21	18	**22**	22	19	**23**	23	20	**24**
5	17	15	**21**	18	16	**22**	20	17	**22**
n	53		**95**	49		**111**	51		**85**

Conclusions

Lithuanian norms for the Raven's *Coloured Progressive Matrices* were developed using data collected from on a representative sample 6 to 11 year-olds. These norms differ from their British and U.S. equivalents.

The study reconfirmed the reliability and validity of the CPM. The norms that have been established can therefore be used with confidence to evaluate the non-verbal reasoning ability of Lithuanian children in the course of formal assessments or screenings. Children living in rural areas and having less educated parents in general scored less well on the CPM. Gender differences were not significant.

References

Aiken, L. R. (2003). *Psychological Testing and Assessment*. Boston: Pearson Education Group.

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. New York: Cambridge University Press.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measures. *Psychological Bulletin, 101*, 71-191.

Gintiliene, G., & Girdzijauskiene, S. (2000). Gender difference in intelligence of Lithuanian children. 11th European Conference on Personality. Freidrich-Schiller – Universitat Jena, 2000, July. *Conference Program and Abstracts* (p.105). *Lengerich: Pabst. Science Publishers*.

Gintiliene, G., & Girdzijauskiene, S. (2003). Lithuania: Culture and children's intelligence. In J. Georgas, L.G. Weiss, J. R. Van de Vijver, & D. H. Saklofske (Eds.), *Cross-cultural Analysis of the WISC-III* (pp.165-179). San Diego: Academic Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press.

Lynn R., & Kazlauskaite V. (2002). The study of IQ in Lithuania. *Perceptual and Motor Skills, 95(2),* 6111- 6112.

Neisser, U. (Ed.). (1998). The Rising Curve. Washington, DC: American Psychological Association.

Raven, J. (1981). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No. 1: The 1979 British Standardisation of the Standard Progressive Matrices and Mill Hill Vocabulary Scales, Together With Comparative Data From Earlier Studies in the UK, US, Canada, Germany and Ireland*. San Antonio, TX: Harcourt Assessment.

Raven, J. (2000a). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No. 3 (Second Edition): A Compendium of International and North American Normative and Validity Studies Together with a Review of the Use of the RPM in Neuropsychological Assessment by Court, Drebing, & Hughes*. San Antonio, TX: Harcourt Assessment.

Raven, J. (2000b). The Raven Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*,1-48.

Raven. J., Raven J. C., Court J. H. (1998a). *Manual for Raven's Progressive Matrices and Vocabuliary Scales, Section 2: Coloured Progressive Matrices*. San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (1998b, updated 2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1: General Overview*. San Antonio, TX: Harcourt. Assessment.

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence*, Volume 2, (pp. 47-103). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chapter 12

The Standard Progressive Matrices in Turkey

N. Ekrem Duzen, Nail Sahin, John Raven, C. Jean Raven

Background

This chapter summarizes the Turkish standardization of the Standard Progressive Matrices (SPM).

The only tests of intelligence or general mental ability in Turkey prior to SPM were the Wechsler Intelligence Scale for Children-Revised (WISC-R) and Stanford-Binet Scale of Intelligence. Both have been widely used in Turkey, especially in health institutions and schools. A Turkish adaptation and standardization of WISC-R had been realized in eighties (Savasir and Sahin, 1988). Although no broadly based adaptation of the Stanford-Binet Scale has been undertaken, clinicians rely on their accumulated experience in interpreting the scores and they use it as a signaling apparatus rather than a measurement device. As a result, the absence of national norms does not appear to be a practical disadvantage although the test increasingly fails to meet the needs of the clinicians with regard to ever increasing complexity of multi-channel information and the test itself has not been brought up to date.

The standardization of the Standard Progressive Matrices (SPM) was begun as part of a larger project concerned with the selection and identification of gifted children. It was the first project of its kind in Turkey. The overall project, consisting of several studies, encompassed a general framework as well as a specific application. An outline of the selection and the identification process for gifted children, including the use of the SPM, can be found in Sahin and Duzen (1994a, 1994b). The more specific Turkish standardization of SPM was summarized by Sahin and Duzen (1994c). Since that time, however, progress has been interrupted due to financial and administrative reasons which were more common in Turkey at that time than they are today.

Sahin and Duzen (1994c) provide preliminary data on the correlations between SPM raw scores and age alongside the correlations between SPM and WISC-R scores (both of Turkish versions). Although these correlation coefficients were derived from a group of potentially high-ability children, they nevertheless indicate something about the reliability of the SPM since the reliability of the WISC-R is well known. Strikingly, these preliminary results were almost identical to those reported by Raven, Court and Raven (1995) for British samples.

Since 1994 no significant contribution to the measurement intelligence and/or general mental ability has been reported for Turkey, although a number of studies were carried out within scope of graduate theses, dissertations, and other academic research projects. However, the need for a non-verbal intelligence test with group administration has been continuously felt not only for studies of specific groups like the gifted but also in connection with such things as vocational training among different age and ability groups. It is therefore with more than a sense of relief that we have at last been able to revisit the original data and produce Turkish norms for the SPM. We hope that, despite their limitations, they will prove useful to professionals and researchers.

The Normative Sample – and Some Information about Turkey.

The size and the structure of the sample on which the data to be presented in this chapter are based was chiefly determined by the design of the original project that incidentally made the SPM standardization possible. The main limitation is that the sample had to be drawn from Ankara, the capital of Turkey. This restriction is not, however, so serious as might appear because of the extensive population movements that have been taking place since the mid 1900s.

Until the seventies, Turkey remained mainly a rural society, carrying with it associated behaviors and beliefs. Although the duration of formal education in the urban areas was not high (3 to 8 years, on the average), in rural parts it was so low (rarely more than three years, even if one had a chance to go to any kind of school) that literacy was a significant index of social status, prestige, and wealth. At the turn of the nineties (ie when the data on which the present study is based were collected), the urban/rural ratio of the country's population had been reversed from 30/70 to 70/30 in less than 30 years. The first decade of the 21st century has brought the ratio to 80/20.

Such vast population shifts from rural to urban areas had changed the country's demography enormously (Basaran, 2004; Baydar, 1999). Most affected were the big cities, the so-called metropolitan areas. More than two thirds of all domestic migration was absorbed by the three biggest cities. The newly settled rural population did not become urbanized easily and quickly. Quite the contrary, adjustment to urban life was, and is, a relatively slow process (Tümertekin, 1968). Therefore, the distinctions between urban and rural were preserved physically and conceptually for long periods in and around urban areas.

Domestic migration is a continuing process, although it might be said that it has slowed down over the last five years. On the other hand, population movements inside the cities have complicated the picture. In other words, since the population flow was too fast, cities had to expand rapidly. In doing so, they did not always offer commensurate expansion of the facilities that a city is expected to offer. The result was a stratification of settlements inside big cities. These represent the urban, the rural, and the transitions between the two (see Benedict, Tümertekin, & Mansur, 1974).

One of the most dramatic indicators of stratification is literacy. The literacy rate hardly reached 70% for males and 50% for females in many rural parts of the country during late eighties and early nineties. There were significant regional differences, with western and northern parts being highly literate whereas middle, eastern, and southeastern parts were highly disadvantaged in terms of literacy and many other factors like economic wealth and liberal social life. These regional differences were directly transferred to urban areas, albeit disproportionately. These transfers caused cities to encompass strata with concomitant literacy ratios.

One particularly striking aspect of this stratification is that many people from certain regions of the country occupy nearby districts in metropols. In other words, it is more or less possible within the big cities to find areas that replicate the demographics of the regions from which their residents came. Although no systematic mapping has been reported, bits and pieces of information make it a well known fact that big cities are composed of small villages, towns, and cities rather than being a unified amalgamation of many elements (Keyder, Aksit, & Aricanli, 1980).

Although there are considerable individual differences within all groups in willingness to adjust to and integrate with urban life, it is very interesting to observe (and it would be challenging to investigate) the way

in which these attitudes and behaviour vary with the area of the country from which they came (Kagitcibasi, 1998).

Given the stratified structure of big cities with respect to regional (including urban/rural) distinctions, it seemed likely that it would be possible to draw a sample that would make it possible to represent such differences. Ankara is the second biggest city in Turkey and had a population of three million at the time of this study. Therefore, it seemed not too difficult to draw a sample which would be reasonably representative of the country as a whole if the stratification was done properly. For this reason, the locations of the schools in the sample were chosen to represent a cross section of the different kinds of district.

In all, 15 primary schools were selected using information supplied by the Ministry of National Education. These schools served different districts of Ankara, ranging from downtown to the extreme periphery. As will be seen later, while most were state schools, some were private.

Within schools, pupils came from all grades from 1st to 8th. A total of 2458 pupils (1170 girls and 1288 boys) were tested. Table 12.1 gives a general picture of the sample.

At this point, in view of its importance later, it may be noted that the urban-rural backgrounds of the pupils was inferred from the location of the schools in which they were enrolled. This was possible because, for example, people from rural areas mostly settle in new areas on the outskirts of the city where people from their hometowns, mostly their relatives (having bonds of kinship in varied degrees) are already living whereas, as indicated above, those living in the city centers mostly come from urban backgrounds.

The previously mentioned information was supplemented by information obtained by questionnaire. For example, information about backgrounds was checked using such information as the names of the

Table 12.1. *Classic Standard Progressive Matrices*
Composition of the Turkish standardization sample.

Grade	1	2	3	4	5	6	7	8	Total
Girls	72	93	145	203	364	152	72	69	**1170**
Boys	64	121	153	253	375	164	87	71	**1288**
Total	136	214	298	456	739	316	159	140	**2458**

areas in which the family had been living and how long they had lived there. Information was also obtained about the public or private schooling of the pupils' parents.

In all, 410 pupils came from private schools and the rest, 2048, from state schools. This ratio is fairly representative of public to private schooling at the time at which the data were collected. Table 12.2 depicts this breakdown.

It may be noted here that private schools also enrolled pupils from rural regions. However, these were mostly the children of people who are 2nd generation of migrants, already adjusted to urban life, rather than new migrants. Such is an increasing trend and, at present, it is possible to observe the 3rd generation of migrants who have almost no affiliation with their rural roots at all.

Below is a summary of additional information about the sample. Table 12.3 depicts the educational and occupational statuses of parents with additional information about the type of place in which they spent most of their lives.

Procedure

Data collection began in the spring of 1992 and finished in spring 1993. Data from the 4th and 5th grades were obtained first. Then, in fall 1992, data from 3rd and 6th grades were collected. Finally, the 2nd and 7th grades and the 1st and 8th grades were tested in spring 1993. All data were collected under direct administration and supervision of the first author. All administrations were on a group basis in the regular classes and classrooms of the schools. All pupils in the selected school classes were tested without further sampling within classrooms or within schools. In order to prevent the pupils making extra preparations, testing was done on normal school days, without prior notice being given to the pupils.

Turkish Norms

Given the limitations of the sample among the younger and older age groups it was decided to produce norms for the age group 6 years 6 months to 14 years 6 months. This reduced the effective sample size from 2458 to 2397. Table 12.4 presents 1993 Turkish norms, which are then compared with the 1979 UK data in Table 12.5.

From Table 12.5 it is clear that the UK norms are, on the whole, slightly higher than the corresponding Turkish figures. Expectedly, the differences are smaller at higher percentiles and larger at lower ones.

Table 12.2. *Classic Standard Progressive Matrices*
Urban/rural background by whether attending a public or private school.

	Public Schools	Private Schools	Total	%
Rural Origins With Different Regional Backgrounds	1341	0	1341	55%
Urban Origins Relatively Homogeneous Backgrounds	707	410	1117	45%

Table 12.3. Parents' educational and occupational status together with area of residence.

PARENTS' EDUCATIONAL STATUS	**MOTHERS**	**FATHERS**
No Schooling	11.2	3.1
Some Primary School (1st to 5th grades)	38.4	27.1
Secondary School (6th to 8th grades)	11.2	16.6
High School or Lyceé (9th to 11th grades)	17.6	18.6
College (2 years) or Higher (4 years) Education	18.4	28.6
Postgraduate degree	3.1	5.8

PARENTS' OCCUPATIONAL STATUS	**MOTHERS**	**FATHERS**
Retired	0.7	3.3
Unemployed (Housewife in case of Mothers)	70.0	1.1
Worker (Blue collar)	2.2	15.7
Worker (White collar government officials)	17.7	27.9
Farmer	0.1	0.7
Craftsman/Shopkeeper	0.8	18.6
Small Business/Free-lance	3.7	18.3
Higher Government Official or Equivalent	3.5	9.1
Private Sector (Own or Higher Manager)	1.3	5.3

LOCATION where pupils spent most of their lives	%
Village (Pop. under 2000)	6.0
Town (Pop. 2000-20.000)	1.6
City (Pop. 20.000-500.000)	23.6
Big City (Pop. over 500.000)	68.9

Table 12.4. *Classic Standard Progressive Matrices*
Smoothed Summary Norms for Turkey (Ankara-stratified) 1993.

	Age in Years (Months)								
	6½	7	7½	8	8½	9	9½	10	10½
Percentile	6(3) to 6(8)	6(9) to 7(2)	7(3) to 7(8)	7(9) to 8(2)	8(3) to 8(8)	8(9) to 9(2)	9(3) to 9(8)	9(9) to 10(2)	10(3) to 10(8)
95	24	29	33	37	42	45	47	47	48
90	20	24	29	34	39	42	44	45	46
75	17	20	25	29	33	37	39	40	41
50	**13**	**15**	**18**	**21**	**24**	**27**	**29**	**31**	**32**
25	10	13	15	17	20	22	24	25	26
10	8	10	11	12	12	13	13	14	14
5	6	8	10	11	11	11	11	12	12
n	67	49	87	104	54	186	108	381	411

(continued)

	Age in Years (Months)							
	11	11½	12	12½	13	13½	14	14½
Percentile	10(9) to 11(2)	11(3) to 11(8)	11(9) to 12(2)	12(3) to 12(8)	12(9) to 13(2)	13(3) to 13(8)	13(9) to 14(2)	14(3) to 14(8)
95	48	48	49	51	52	52	52	53
90	46	46	47	49	51	51	51	52
75	42	42	42	43	44	46	48	50
50	**33**	**34**	**34**	**35**	**36**	**39**	**41**	**43**
25	27	27	28	28	28	28	29	32
10	14	14	14	15	15	17	18	19
5	12	12	12	12	12	12	13	14
n	274	115	168	104	119	59	72	38

Contributions to the Variance in Scores

Given what has been said about the relatively recent emergence of Turkey from being a predominantly rural society, it seemed important to try to assess the potential importance of this factor in creating the UK-Turkish difference. Our first step toward doing this was to calculate the mean scores for each age group for the urban and ex-rural populations separately. The results are presented in Table 12.6. This shows that the scores of the pupils studying in schools serving areas in which the majority of the population had been living in towns for some time are significantly ($p<.001$) higher than those of pupils living in schools serving mainly populations who had recently migrated from rural areas. It should be noted that urban-

Table 12.5. *Classic Standard Progressive Matrices*
1993 Turkish data in the context of 1979 British norms.

	Age in Years (Months)																
	6½ 6(3) to 6(8)		7 6(9) to 7(2)		7½ 7(3) to 7(8)		8 7(9) to 8(2)		8½ 8(3) to 8(8)		9 8(9) to 9(2)		9½ 9(3) to 9(8)		10 9(9) to 10(2)		10½ 10(3) to 10(8)
Percentile	UK	**TR**	UK	**TR**	UK	**TR**	UK	**TR**	UK	**TR**	UK	**TR**	UK	**TR**	UK	**TR**	UK
95	33	**24**	34	**29**	37	**33**	40	**37**	42	**42**	44	**45**	46	**47**	48	**47**	49
90	30	**20**	32	**24**	35	**29**	38	**34**	40	**39**	42	**42**	44	**44**	46	**45**	47
75	22	**17**	26	**20**	30	**25**	33	**29**	36	**33**	38	**37**	41	**39**	42	**40**	43
50	**16**	**13**	**19**	**15**	**22**	**18**	**25**	**21**	**31**	**24**	**33**	**27**	**36**	**29**	**38**	**31**	**39**
25	13	**10**	14	**13**	15	**15**	17	**17**	22	**20**	25	**22**	28	**24**	32	**25**	33
10	10	**8**	12	**10**	12	**11**	14	**12**	17	**12**	17	**13**	19	**13**	23	**14**	27
5	9	**6**	10	**8**	11	**10**	12	**11**	13	**11**	14	**11**	14	**11**	17	**12**	22
n	112	**67**	138	**49**	148	**87**	174	**104**	153	**54**	166	**186**	198	**108**	172	**381**	194

(continued)

	Age in Years (Months)																
	11 10(9) to 11(2)		11½ 11(3) to 11(8)		12 11(9) to 12(2)		12½ 12(3) to 12(8)		13 12(9) to 13(2)		13½ 13(3) to 13(8)		14 13(9) to 14(2)		14½ 14(3) to 14(8)		15 14(9) to 15(2)
Percentile	UK	**TR**	UK	**TR**	UK	**TR**	UK	**TR**	UK	**TR**	UK	**TR**	UK	**TR**	UK	**TR**	UK
95	50	**48**	51	**48**	52	**49**	53	**51**	54	**52**	54	**52**	55	**52**	56	**53**	57
90	48	**46**	49	**46**	50	**47**	51	**49**	52	**51**	53	**51**	54	**51**	54	**52**	55
75	44	**42**	45	**42**	46	**42**	47	**43**	49	**44**	49	**46**	50	**48**	50	**50**	51
50	**40**	**33**	**41**	**34**	**41**	**34**	**42**	**35**	**43**	**36**	**44**	**39**	**45**	**41**	**46**	**43**	**47**
25	34	**27**	36	**27**	37	**28**	38	**28**	39	**28**	41	**28**	42	**29**	42	**32**	42
10	29	**14**	31	**14**	31	**14**	32	**15**	33	**15**	35	**17**	36	**18**	36	**19**	36
5	24	**12**	25	**12**	26	**12**	27	**12**	28	**12**	29	**12**	30	**13**	33	**14**	33
n	187	**274**	164	**115**	164	**168**	174	**104**	185	**119**	180	**59**	196	**72**	189	**38**	191

rural differences may have been contributing significantly to the obvious differences between higher and lower percentiles. As numerous studies reviewed in the Manual for Raven's Progressive Matrices and Vocabulary Scales (Raven, Raven, & Court, 1998 [updated 2003]; 2000 [updated 2004]; Court & Raven (1995) – and especially the Irish standardization (Raven, 1981) – have shown, major urban-rural differences are to be expected. (While there were major differences between different areas of the UK at the time of the 1979 standardization [with the Monklands area of Scotland, which then had the worst and most distinctive socio-economic conditions in Europe, and not the more rural areas, having the lowest scores] these differences were entirely explained by variance in socio-economic status.)

Table 12.6. *Classic Standard Progressive Matrices*
Mean scores of pupils studying in schools primarily serving long standing urban populations compared with those primarily serving populations recently migrated from rural areas.

	SPM Total Score			
	Urban		Rural	
AGE	Mean	N	Mean	N
6½	14.9	45	14.7	22
7	15.1	34	12.9	15
7½	23.7	68	17.2	19
8	23.2	51	15.4	55
8½	33.9	35	25.7	18
9	36.7	42	20.9	144
9½	33.9	58	26.8	50
10	38.5	118	24.7	263
10½	38.1	307	30.8	104
11	36.8	140	25.4	134
11½	33.3	30	29.9	85
12	33.9	66	27.6	102
12½	38.1	31	32.9	73
13	38.9	40	34.5	79
13½	44.3	13	39.0	46
14	46.5	15	34.5	57
14½	46.1	12	38.3	26

Having shown that scores do vary with whether pupils came from predominantly an urban vs. a rural background and noted that, at least in the UK and USA, much of the variance between areas was accounted for by variance in socio-economic status, it seemed important to investigate the relative importance of urban-rural origins and socio-economic status in Turkey.

Socio economic status is generally assessed using an index based on the income and prestige of the main breadwinner's occupation. However, as has also proved to be the case with the concept of "intelligence", while many variables – such as cultural level of the home, childrearing practices, and parental values – are associated with this basic dimension, attempts to unscramble their differential effects have proved notably unsuccessful. In contrast, a single index of the social prestige of the main breadwinner's

occupation has turned out to be remarkably robust. Unfortunately, as is the case in the present study, it has not always proved easy to get reliable information on parental occupation or income, while data on parental education seems easier to collect.

It cannot, however, necessarily be assumed that the relationships summarized in the last paragraph will apply in Turkey. It has, for example, been observed that, while higher levels of income contribute to the acquisition of higher educational qualifications, the converse less often follows: higher educational qualifications do not necessarily lead to higher income. Equally, the educational level of the mother may be more important than that of the father.

In an attempt to assess the relative importance of each potentially relevant variable, multiple regression techniques were employed. The first step was, as in Table 12.7, to partial out the effects of age.

The next finding was that, once the effects of age had been removed, as hinted above, the data on the main breadwinner's occupation was so unreliable that it failed to account for any of the remaining variance in SPM scores. That left urban vs. rural background, mother's education, and father's education.

Across the whole sample, the correlation between age and RPM score was .38, implying that it accounted for 14% of the total variance. The combined predictive power of all the variables together yielded a correlation of .55, or just over 30% of the variance.

Age and mother's education level together yielded a multiple correlation of .52. Age, mother's education, and urban-rural origin upped this to .54. Thus it seemed that mother's education accounted for most of the variance attributed to urban-rural origin in Table 12.6. Finally, the addition of Father's occupation raised the multiple correlation coefficient to the .55 mentioned above.

In an effort to clarify whether mother's education or rural vs urban living was the more basic variable, the regression analysis was re-run forcing in rural-urban first (Table 12.8). Even so, the rural-urban distinction could explain only a small portion of the variance (R^2=.075), while rural-urban and mother's education together explained more (R^2=.124). When father's education was added R^2 rose to .133. When the order of variables was further manipulated, urban-rural being the first and mother's education the last, the picture remained much the same.

These results clearly indicate that, although the differences between life in rural and urban areas do have a lasting effect on test scores, that effect is trivial compared with the effects of mother's level of education.

Table 12.7. *Classic Standard Progressive Matrices*
Correlations between variables contributing to variance in scores

	SPM Total Score	Age	Mother's Education	Rural-Urban
Age	.38**			
Mother's Education	.28**	-.21**		
Rural-Urban	.28**	-.17**	.67**	
Father's Education	.25**	-.19**	.77**	.61**

** Correlation is significant at the 0.01 level (2-tailed).

Table 12.8. *Classic Standard Progressive Matrices*
Correlations between variables contributing to variance in SPM score, with grade substituted for age

	SPM Total Score	Grade	Mother Education	Rural-Urban
Grade	.46**			
Mother's Education	.28**	-.14**		
Rural-Urban	.28**	-.08**	.67**	
Father's Education	.25**	-.12**	.77**	.61**

To put it differently, urban-rural backgrounds take effect only if mothers' educational level is in expected direction. Then, it may be argued that mothers might be attaining adequate levels of education to offset disadvantages arising from rural backgrounds. However, the case is somewhat otherwise. The level of education that mothers in urban areas might attain is quite low, on the average, and it was much lower in the early nineties. This may sound a bit counterintuitive but given the fact that, except for their historically urbanized portions, the big cities in Turkey are indeed just recently urbanized settlement areas which are mainly inhabited by people of rural origins. Although such a mode of urbanization might lead to the attainment of some higher level of education, it did not create a real distinction from actual rural parts of the country. Therefore, similarity in background might be sufficient to lead pupils of literate mothers from actual rural backgrounds to surpass their cohorts who were children of mothers who were living in cities but with somewhat 'concealed' rural origins with inadequate education levels.

In the light of these results, it seemed worth checking on another variable that might help to explain the non-age variance. The generic inference here is that non-age variance could be accounted by factors which are in some way compensating for simple age differences. A meaningful

test of this observation is to try the same analyses with another variable which could be used as equivalent to age. In our case, there appeared another variable which could be used to indicate the developmental differences as reliably as age groups but with more predictive value. That variable is school grade, in the sense of "year-group within the school system" rather than "mark".

Although the correlation between age and grade is .94 ($p < .0001$), grade predicts SPM score more accurately than does age. The raw correlation is .46 (as against .38). When grade is combined with mother's educational level, the multiple correlation increases to .57. And grade, mother's education, and urban-rural origin bring the figure up to .58. Finally, it reaches the highest level when all variables (class, mother's educational level, urban-rural origin, and father's educational level) are included ($R = .59$). It is once more apparent that, even though father's educational level does indeed make a contribution to "explaining" the variance, it adds little on its own.

Distributions of SPM Total Scores by Age Group

Figure 12.1. Classic Standard Progressive Matrices
Turkish Standardisation
Distribution of Total Scores by Age Group

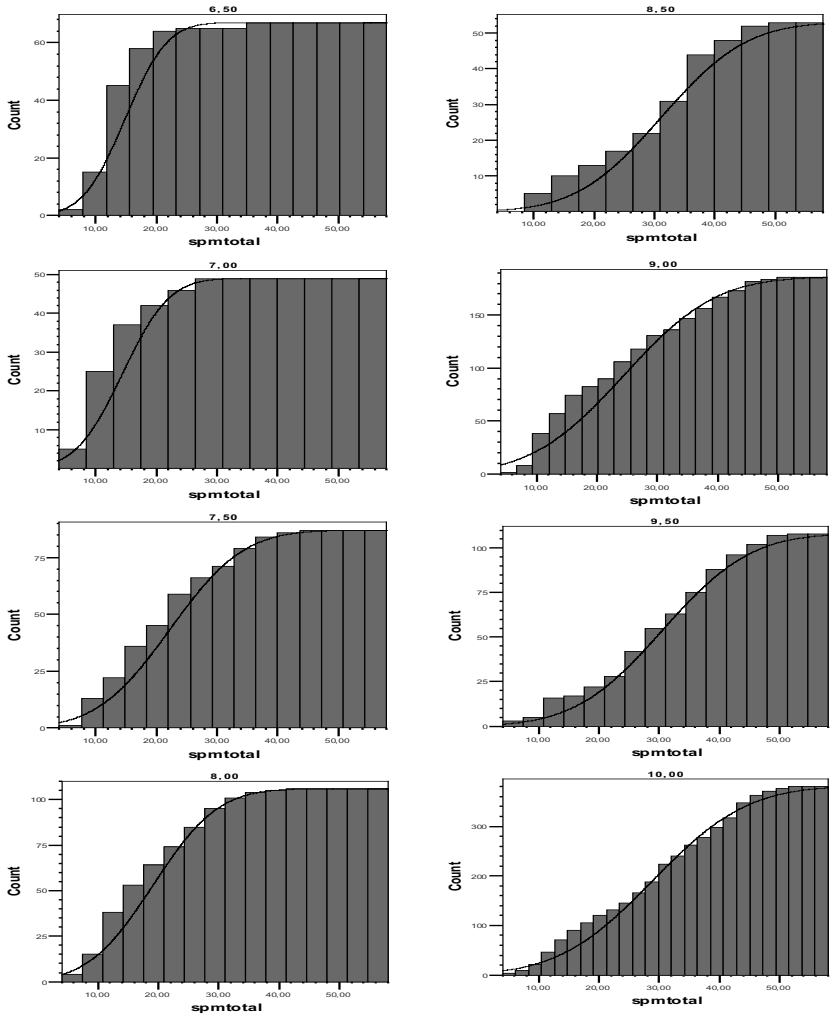
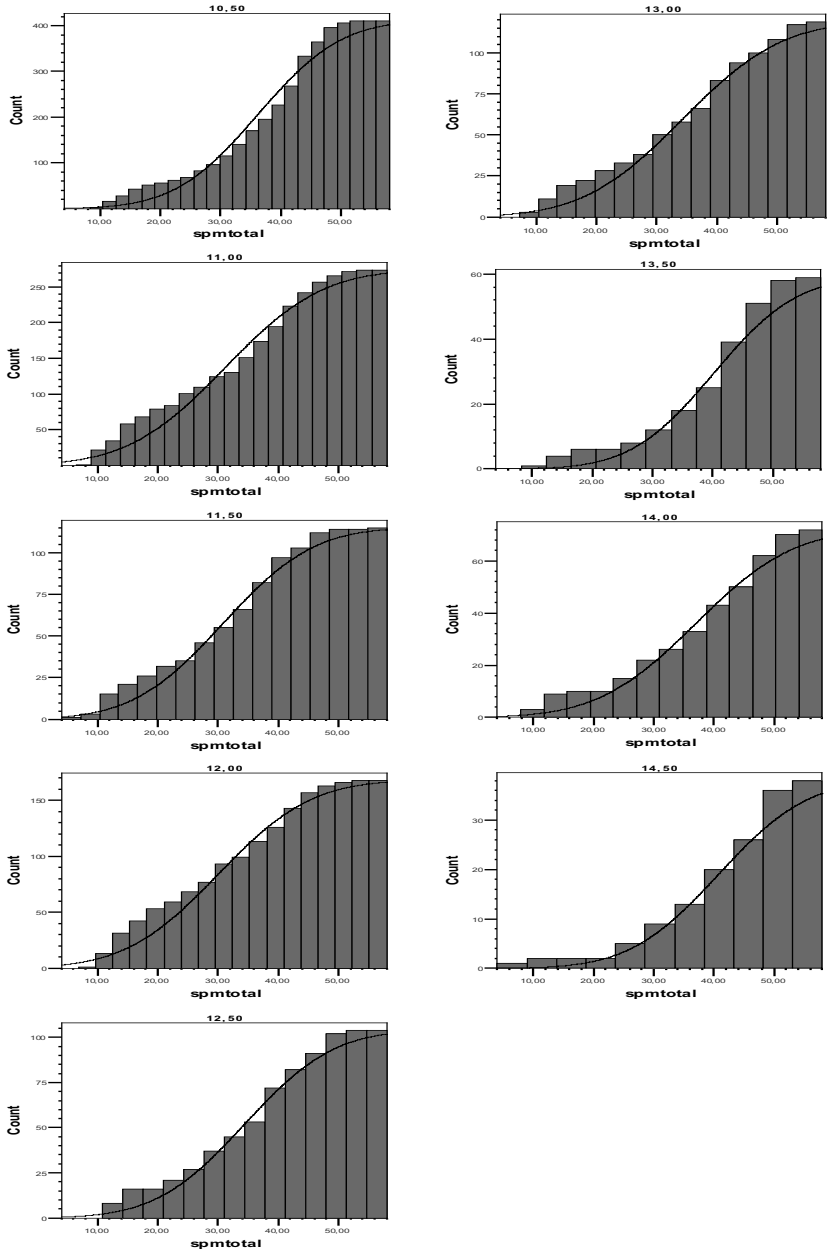


Figure 12.1. Classic Standard Progressive Matrices
Turkish Standardisation
Distribution of Total Scores by Age Group
(continued)



Conclusions

The Turkish standardization of the SPM unexpectedly shed light on a number of points which had not previously been part of our concerns. The results lead us to search for new variables to better explain non-age variance. One noticeable result of this was that urban-rural background, as such, turned out to have much less effect on pupils' performance than expected. More important was the educational level of the parents – and especially that of the mother. This result to a degree calls into question the widespread belief that a rural background by itself will have deleterious effects due to lack of countless facilities and conveniences associated with urban life. It is not immediately obvious that the effects of such disadvantages can be easily counteracted by such a "simple" thing as mother's level of education.

It is possible to speculate more on the hows and whys of the effects of mother's education given the stratified nature of both urban and rural backgrounds specific to Turkey. However, to argue on the basis of the data at hand, we might refer to another variable which also functions to explain non-age variance. This variable is grade. It is interesting to observe that the correlations between the demographic variables (especially urban-rural) and grade are so much lower than the equivalent correlations with age. Moreover, the correlation between SPM Total Score and Grade is .46 – in contrast to a correlation of .38 with age. It may be that grade is functioning as some kind of compensatory variable to offset the effects of other demographic variables, including the urban-rural distinction, which are expected to contribute to differences in scores. However, in our case, grade seems to steal those effects from other variables. Why grade rather than age brings about such a balance could be explained on the basis of its being a factor of adjustment. The most likely explanation could be that 'grade' here might be playing a socializing role (with all motivational attributes attached to it) something which a more neutral factor like age cannot operate likewise. Consequently, when pupils of different ages gather around the same grade they enjoy the benefits of the contacts that that network can bring about. This is to say, grade provide pupils with a way to meet their intrinsic need to compensate the disadvantages of different (or not commonly shared) backgrounds.

In this sense, the power of mother's educational level over non-age variance gains more meaning. This explanation points to the conclusion that when and where pupils find a chance to balance out effects arising

from their disadvantageous backgrounds they immediately use it as much as they can, whether this be their mothers' or cohorts' educational levels, or something else.

The features of the standardization sample and the results summarized in this chapter allow us to suggest that the urban parts of Turkey are indeed not 'urban' enough to justify producing different norms for urban and rural regions or backgrounds. The critical point here is that urban backgrounds are not readily associated with access or attainment to higher education levels. Rather, the two are much closer than it might first seem that even a slight difference originating from a positive contribution to disadvantageous backgrounds balances such expected differences.

It might be interesting to investigate the longitudinal alterations and developments in mutual positions of different regional backgrounds around demographic variables as expressed in test performances. Subsequently, follow up studies with SPM and further measurements with other types of Progressive Matrices, including Advanced Progressive Matrices, have the potential to reveal the nature and sources of individual differences. In this sense, measurement of intelligence could once more be conceived an issue of a broader understanding than an issue of assessing psychometric properties of a mental performance.

References

Basaran, F. (2004). *Gecis Doneminde Turkiye: Degisim, Gelisim, Tutumlar ve Degerler*. Ankara: Türk Psikologlar Dernegi Yayınları.

Baydar, O. (1999). *75 Yilda Köylerden Sehirlere*. Istanbul: Tarih Vakfı Yurt Yayınları.

Benedict, P., Tümertekin, E., & Mansur, F. (1974-Eds.). Turkey: geographic and social perspectives. *Social, Economic, and Political Studies of the Middle East, v. 9*. Leiden: Brill.

Court, J. H., & Raven, J. (1995). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 7: Research and References: Summaries of Normative, Reliability, and Validity Studies and References to All Sections*. San Antonio, TX: Harcourt Assessment.

Kagitcibasi, C. (1998). *Kulturel Psikoloji: Kultur Baglaminda Insan ve Aile*. Istanbul: Yapi Kredi Yayinlari.

Keyder, C., Aksit, B., & Aricanli T. (1980). *Paths of rural transformations in Turkey*. Institute of Economical and Social Research, Working Paper, 11. Istanbul: Bosphorus University.

Raven, J. (1981). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.1: The 1979 British Standardisation of the Standard*

Progressive Matrices and Mill Hill Vocabulary Scales, Together With Comparative Data From Earlier Studies in the UK, US, Canada, Germany and Ireland. San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (1998, updated 2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1*: *General Overview*. San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3*: *The Standard Progressive Matrices,I including the Parallel and Plus Versions*. San Antonio, TX: Harcourt Assessment.

Sahin, N., & Duzen, N. E. (1994a). A multi-step selection process for the high ability children. In K.A. Heller and E.A. Hany (Eds.), *Competence and Responsibility*: *Proceedings of the 3rd ECHA Conference, Vol. 2*, 280-285. Göttingen: Hogrefe & Huber Publishers.

Sahin, N., & Duzen, N. E.(1994b). The gifted child stereotype among university students and elementary school teachers. In K.A. Heller and E.A. Hany (Eds.), *Competence and Responsibility*: *Proceedings of the 3rd ECHA Conference, Vol. 2*, 367-376. Göttingen: Hogrefe & Huber Publishers.

Sahin, N., & Duzen, N. E. (1994c). *Turkish standardization of Raven's Standard Progressive Matrices Test in 7-15 year-old Turkish children*. Paper presented at the 23rd International Congress of Applied Psychology, July 17-22, Madrid, Spain.

Savasir, I., & Sahin, N. (1988). Wechsler Cocuklar icin Zeka Olcegi Turkiye Uyarlaması. Ankara: Türk Psikologlar Dernegi Yayınları.

Tümertekin, E. (1968). *Internal Migration in Turkey*. Istanbul: Publications of Istanbul University (# 1371).

Chapter 13

Kuwaiti norms for the Classic SPM in an International Context*

Ahmed Abdel-Khalek** and John Raven

Abstract

A probability sample (*n*=8,410) of Kuwaiti school students aged 8-15 responded to Raven's *Standard Progressive Matrices*. The test was administered, untimed, in group sessions. In this paper, the smoothed summary age norms for Kuwait (which will themselves be of interest to many psychologists and others working in Kuwait and neighbouring countries) are first compared with what have become the standard international reference data for such work, namely the 1979 British norms. Thereafter, the Kuwait norms are compared with those reported for several other countries and set in the context of accumulating data on changes over time. The results show that, while some, as yet unidentified, features of the environment do have a dramatic effect on scores, aspects of the environment that many people would have expected to have a significant effect (such as differences in calligraphy) are much less important than might have been thought.

Background

Raven's Standard Progressive Matrices (SPM) test was constructed to measure the eductive component of ***g*** as defined in Spearman's theory of cognitive ability (Raven, Raven, & Court, 1998, updated 2003, p. G1).

* An earlier version of this chapter was published in *Social Behaviour and Personality: An International Journal* (2006), *4,* 169-179.

** The research reported in this chapter was supported by Kuwait University under Grant No. OP01/02. The authors gratefully acknowledge the able assistance of Research Administration at that university.

Kaplan and Saccuzzo (1997) stated that "research supports the Raven Progressive Matrices (RPM) as a measure of general intelligence, or Spearman's ***g*** factor… In fact, the Raven may be the best single measure of ***g*** available" (p. 359).

In the same vein, Jensen (1998) maintained that "in numerous factor analyses, the Raven tests, when compared with many others, have the highest ***g*** loading and the lowest loadings on any of the group factors. The total variance of Raven scores in fact comprises virtually nothing besides ***g*** and random measurement error" (p. 541). He added that Raven's Progressive Matrices is often used as a "marker test of Spearman's ***g***. That is, if it is entered into a factor analysis with other tests of unknown factor composition, and if the Matrices has a high loading on the general factor of the matrix of unknown tests, its ***g*** loading serves as a standard by which the ***g*** loadings of the other tests in the battery can be evaluated" (p. 38).

By the same token, Lynn, Allik, Pullman, and Laidra (2004) stated that "the Progressive Matrices is widely regarded as the best test of abstract or non-verbal reasoning ability, and this is itself widely regarded as the essence of "fluid intelligence" and of Spearman's ***g***" (p. 1250). Mackintosh (1996, p. 564) has described it as "the paradigm test of non-verbal, abstract reasoning ability".

This view is not, of course, universally accepted. Indeed, Raven, Raven and Court (1998, 2000) refer to several studies which suggest a loading on spatial ability, and a review of the extensive literature dealing with this topic from the point of view of researchers keen to distinguish "Working Memory" from ***g*** has been provided by Ackerman, Beier, and Boyle (2002).

The Standard Progressive Matrices test enjoys good psychometric characteristics (see: Court & Raven, 1995; Kline, 2000; Murphy & Davidshoffer, 1998). A huge body of published research bears on the validity of this test (Gregory, 1992). Therefore, it has gained widespread acceptance and use in many countries all over the five continents (Irvine & Berry, 1988). No other test has been so extensively used in cross-cultural studies of intelligence. Lynn and Vanhanen (2002) summarized a plentiful number of studies based on normative data for the test has been collected in 61 countries. For all these reasons, Kaplan and Saccuzzo (1997) concluded that "with its new worldwide norms and updated test manual, the Raven holds promise as one of the major player in the testing field in the 21st century" (p. 361).

The Arab countries are in a great need of standardized intelligence tests with local norms. Indeed, the three series of the Progressive Matrices test, i.e., the Standard, Coloured, and Advanced, are already available in the majority of the Arab countries. The Standard Progressive Matrices test has been administered to different samples in most Arab countries. However, the vast majority of these studies remain either unpublished or published in Arabic.

In 1988, Abdel-Khalek found that the test-retest reliability reached .82 among Egyptian college students. A clear general factor with high loadings was extracted from the five Sets of the test. A factor analysis of the total score on the test and four subscales of Thursone's Primary Mental Abilities yielded a general and high loaded factor, on which the Matrices loading was .77, denoting high concurrent validity.

More recently, Abdel-Khalek and Lynn (2006) examined the sex differences on the test and found a small sex difference of .08 *sd* (1.2 IQ points) favouring girls.

Using a Kuwaiti sample of school children (n=968), Abdel-Khalek (2005) found the test-retest reliability ranged from .69 to .85, while internal consistency assessed by the alpha coefficient ranged between .88 to .93 denoting good temporal stability and internal consistency. The loadings of the five sets on the only salient factor ranged from .72 to .89 indicating the good factorial validity of the scale.

The objective of the current investigation was primarily to create Kuwaiti norms for the *Standard Progressive Matrices* (SPM) test, but these norms are presented here in an international context since the comparative data that have emerged are of considerable importance to cognitive psychology.

Method

Participants

A sample of 8,410 8-15 year olds was recruited. All of them were Kuwaiti citizens and students in the governmental schools in the six districts in Kuwait. In each district one elementary, intermediate and secondary school for both boys and girls were randomly chosen. The selection of school districts used a stratified random sampling procedure. The test was administered to at least 60 students in each age group of boys and the same for girls in each of the six districts of Kuwait.

The Test

The original, 1958, version of the SPM (Raven, J. C., 1958) was employed but adapted in the sense that, in the Arabic test booklets, the main matrix and the six or eight alternatives were transposed to read from right to left following the custom of Arabic writing.

Procedure

The SPM was administered to students by a group of competent and trained testers. The testers in the boys' schools were male, and female in girls' schools. In every class, testing was carried out by a tester and an assistant. Testing was carried out in whole classes of 25-30 students. Verbal instructions were given to the students on how to do the test. The test was given without time limits. The testing was carried out in the year 2002. The raw data of the completed answer sheets were scored by computer.

Results and Discussion

Table 13.1 presents the Kuwaiti norms in the context of the 1979 norms for Great Britain. It will be seen that, although the scores obtained in Kuwait are, in general, considerably lower than the much earlier UK norms, especially among the younger and less able pupils, the general impression is one of rather surprising similarity. There is not space here to speculate on the reasons for the decline in the discrepancy among the older age groups, but it would be inappropriate not to draw attention to the ceiling effect which now restricts the variation in scores among the more able from age 12 onwards (and which exacerbates the non-Gaussian within-age distributions which, among other things, make it inappropriate to process the data in terms of means and Standard Deviations, let alone IQ scores).

In the earlier version of this paper which was acknowledged on the title page of this chapter, data from an earlier, smaller, sample of Kuwaiti pupils were included in a table comparing a wide range of international norms. Since a modified version of this Table now appears in the *General Introduction* to this book that, rather long, table has been omitted here.

Nevertheless, it would be inappropriate to pass on without comment - especially as the preparation of that table was prompted by the arrival

Table 13.1. *Standard Progressive Matrices*
Smoothed 2006 Norms for Kuwait in the Context of 1979 British Standardisation

	Age in Years (Months)										
	7 6(9) to 7(2)		7½ 7(3) to 7(8)		8 7(9) to 8(2)		8½ 8(3) to 8(8)		9 8(9) to 9(2)		9½ 9(3) to 9(8)
Percentile	UK	**KU**	UK	**KU**	UK	**KU**	UK	**KU**	UK	**KU**	UK
95	34	**29**	37	**33**	40	**38**	42	**40**	44	**42**	46
90	32	**25**	35	**29**	38	**34**	40	**37**	42	**39**	44
75	26	**19**	30	**22**	33	**26**	36	**31**	38	**33**	41
50	**19**	**14**	**22**	**16**	**25**	**18**	**31**	**21**	**33**	**24**	**36**
25	14	**12**	15	**12**	17	**13**	22	**14**	25	**15**	28
10	12	**10**	12	**10**	14	**11**	17	**11**	17	**12**	19
5	10	**9**	11	**9**	12	**10**	13	**10**	14	**10**	14
n	138	**348**	148	**495**	174	**273**	153	**413**	166	**387**	198

(continued)

	Age in Years (Months)										
	10½ 10(3) to 10(8)		11 10(9) to 11(2)		11½ 11(3) to 11(8)		12 11(9) to 12(2)		12½ 12(3) to 12(8)		13 12(9) to 13(2)
Percentile	UK	**KU**	UK	**KU**	UK	**KU**	UK	**KU**	UK	**KU**	UK
95	49	**46**	50	**47**	51	**48**	52	**48**	53	**50**	54
90	47	**43**	48	**44**	49	**46**	50	**46**	51	**48**	52
75	43	**39**	44	**41**	45	**43**	46	**44**	47	**44**	49
50	**39**	**32**	**40**	**35**	**41**	**37**	**41**	**38**	**42**	**39**	**43**
25	33	**23**	34	**27**	36	**30**	37	**32**	38	**33**	39
10	27	**15**	29	**18**	31	**20**	31	**22**	32	**23**	33
5	22	**13**	24	**14**	25	**15**	26	**16**	27	**17**	28
n	194	**405**	187	**436**	164	**418**	164	**401**	174	**395**	185

(continued)

	Age in Years (Months)										
	14 13(9) to 14(2)		14½ 14(3) to 14(8)		15 14(9) to 15(2)		15½ 15(3) to 15(8)		16 15(9) to 16(2)	16½ 16(3) to 16(8)	17 16(9) to 17(2)
Percentile	UK	**KU**	UK	**KU**	UK	**KU**	UK	**KU**	**KU**	**KU**	**KU**
95	55	**52**	56	**54**	57	**54**	57	**54**	**54**	**54**	**54**
90	54	**50**	54	**51**	55	**52**	55	**52**	**52**	**52**	**52**
75	50	**47**	50	**49**	51	**49**	51	**49**	**49**	**49**	**49**
50	**45**	**43**	**46**	**44**	**47**	**45**	**47**	**45**	**45**	**45**	**45**
25	42	**37**	42	**39**	42	**40**	42	**40**	**40**	**41**	**41**
10	36	**31**	36	**32**	36	**33**	36	**33**	**33**	**34**	**34**
5	30	**26**	33	**29**	33	**30**	33	**31**	**31**	**31**	**31**
n	196	**404**	189	**445**	191	**418**	171	**391**	**466**	**425**	**234**

of the Kuwait data. Two things strike one immediately. The first is the similarity between the normative data collected by different people using different sampling procedures in these different countries. The second is the wide within-age variance in the norms from within each country.

The similarity in the norms across countries having such different calligraphies, such different reading and writing systems, such different values, such different educational systems, such different child rearing practices, such different family sizes, such differential access to television, and at such different stages in economic development strongly suggests that cultural variation in these socio-demographic characteristics has much less impact that is commonly assumed.

Furthermore, the variance *within* countries reconfirms this observation. If these cultural variables *did* have the impact on scores that is often asserted they would surely influence the within-culture variance. Everyone in each of these cultures is exposed to much the same cultural environment, yet it seems that it neither restricts nor enhances the within-cultural variance.

From the data for 11 year olds, it would seem that the norms for the 50th and lower percentiles in India, Kuwait, and Qatar lag increasingly behind.

Missing from the table are some data that many people find embarrassing and which lack political correctness. The data in question have to do with Blacks in the USA and South Africa, many Native American groups (with the exception of the Eskimos), and other groups lacking a tradition of literacy.

It would in fact have been misleading to have included these data in the Table because most of the samples leave much to be desired. Nevertheless such data as exist (see Raven, 2000, Court & Raven, 1995, for a summary) taken together with the data for South Africa and Indian tribal areas included in this volume reveal huge differences between these groups and the Kuwait data reported in Table 13.1 as well as that summarised in Table 1.2 of the General Introduction to, and Overview of, this book.

Changes Over Time

It is important to contextualise these observations by again emphasising that data presented in other chapters in this volume reveal dramatic changes in scores over time. This means that some aspects of the environment *do* have an enormous (and previously unsuspected) effect

on scores - but again without significantly reducing the within-age variance. On the other hand, the cross cultural data reported in this book - including that presented in this Chapter - clearly show that the features of the environment that most people first think of as explanations for the change over time have much less effect than was previously believed.

In an effort again to avoid creating a misleading impression it is important to draw attention to the fact that similar increases in scores over time have been documented on a wide range of *verbal* measures of *eductive* (i.e. "meaning making" or "reasoning") ability (see, e.g. Bouvier, 1969; Schaie & Willis, 1986; Flynn, 2000). In other words, they have not been limited to pictorial or diagrammatic tests of the kind we have been concerned with here. This undermines yet another of the "explanations" of the change over time that most commonly spring to mind.

It follows from these observations that the increase is *not* due to such things as schools attempting to enhance levels of "creativity" by encouraging children to tackle non-verbal puzzles or handle computer games or any of the explanations most widely favoured by psychologists and listed by Thorndike (1975, 1977) as possible explanations of the increase he had documented in the norms for the Stanford-Binet test.

From the point of view of seeking an explanation of the increase over time it is, however, perhaps still more important to note that an increase of exactly the same magnitude has occurred in height and life expectancy. It is worth dwelling on some of the implications of this.

First, no one would conclude from the fact that life expectancy is measured by a Rasch Scale analogous to that used to measure eductive ability via the RPM that the variance must be determined by some single underlying ability analogous to speed of neural processing. Nor would they seek a single factor explanation of the increase over time. Nor would they conclude from the fact that backward projection of the increase to the time of the Greeks that the Greeks must have had impossibly short life expectancies and therefore that the measures must be devoid of meaning. Nor would they conclude from the fact that there are ethnic and socio-economic differences in life expectancy that both the measures themselves and the differences between groups have no meaning. And nor would they expect that the same factors as are responsible for the within cohort variance are the same as those responsible for the increase over time. Yet all of these claims have been made by Flynn or others in connection with the increase in Raven Progressive Matrices (RPM) scores.

Finally, it is worth nothing that the causes of the increase in height and life expectancy and the ethnic and cultural differences associated with them have proved just as difficult to pin down as those on the RPM.

Despite these caveats, it is essential, when seeking to interpret the similarities and differences between cultures at any point in time that were summarised in earlier chapters of this volume, to bear in mind that the cross-birth-cohort data show that changes in the environment not only *can*, but *have*, had effects which completely swamp the differences between cultural groups.

Conclusion

Normative data for the Standard Progressive Matrices derived from testing a large representative sample of young people in Kuwait are expected to be of considerable interest to psychologists, teachers, and others working in Kuwait and neighbouring countries. However, when viewed in the context of parallel data from several other countries and cultures, the data acquire a much wider significance in that they reveal remarkable similarity in the norms across cultures at any point in time accompanied by dramatic change over time.

The data clearly show that variation in features of the environment that many people would have expected to markedly influence scores – such as variance in calligraphy, educational systems, and cultural norms – have much less effect than many people would have expected whilst as yet unidentified features of the environment have a much greater effect than many people would have suspected.

In this a context, such cross-cultural differences as remain appear to merit less attention than might otherwise have seemed to have been the case.

And, when the cross-birth-cohort data are themselves compared with similar data relating to life expectancy, the logic of many arguments put forward by psychologists would seem to be, at best, highly questionable.

References

Abdel-Khalek, A. M. (1988). Egyptian results on the Standard Progressive Matrices. *Personality and Individual Differences, 9,* 193-195.

Abdel-Khalek, A. M. (2005). Reliability and factorial validity of the Standard Progressive Matrices among Kuwaiti children ages 8 to 15 years. *Perceptual and Motor Skills, 101,* 409-412.

Abdel-Khalek, A. M., & Lynn, R. (2006). Sex differences on the Standard Progressive Matrices and in educational attainment in Kuwait. *Personality and Individual Differences, 40*, 175-182.

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General, 131,* 567-589.

Bouvier, U. (1969). *Evolution des Cotes a Quelques Test.* Belgium: Centre de Recherches, Forces Armees Belges.

Cayssails, A. (2001). *Carpeta de Evaluacion Escale General*. Buenos Aires, Argentina: Paidos.

Court, J. H., & Raven, J. (1995). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 7: Research and References: Summaries of Normative, Reliability, and Validity Studies and References to All Sections*. San Antonio, TX: Harcourt Assessment

Deltour, J. J. (1993). *Echelle de Vocabulaire Mill Hill de J. C. Raven: Adaptation Francaise et normes comparees du Mill Hill et du Standard Progressive Matrices (PM38). Manuel et Annexes*. Braine le Chateau, Belgium: Editions L'Application des Techniques Modernes SPRL.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29-51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171-191.

Flynn, J. R. (2000). IQ gains, WISC subtests and fluid ***g***: ***g*** theory and the relevance of Spearman's hypothesis to race. In G. R. Bock, J. A. Goode, & K. Webb (Eds.), *The Nature of Intelligence (Novartis Foundation Symposium 233)* pp. 202-227. Chichester, England: Wiley.

Gregory, R. J. (1992). *Psychological testing: History, principles, and applications*. Boston: Ally & Bacon.

Irvine, S. H., & Berry, J. W. (Eds.) (1988) *Human abilities in cultural context*. Cambridge: Cambridge University Press.

Jaworowska, A., & Szustrowa, T. (1991). *Podrêcznik do Testu Matryc Ravena: Wersja dla Zaawansowanych (1992). Polskie standaryzacje: Uczniowie 12;6-19;5 i studenci*. Warsaw: Pracownia Testów Psychologicznych Polskiego Towarzystwa Psychologicznego.

Jensen, A. R. (1998). *The **g** factor: The science of mental ability*. Westport: Praeger.

Kaplan, R. M., & Saccuzzo, D. P. (1997). *Psychological testing: Principles, applications, and issues* (4th ed.). Pacific Grove: Brooks/Cole.

Kline, P. (2000) *Handbook of psychological testing* (2nd ed.). London: Routledge.

Lynn, R., & Vanhanen, T. (2002). *IQ and the wealth of nations*. Westport, CT: Praeger.

Lynn, R., Allik, J., Pullman, H., & Laidra, K. (2004). Sex differences on the progressive matrices among adolescents: Some data from Estonia. *Personality and Individual Differences, 36*, 1249-1255.

Mackintosh, N. J. (1996). Sex differences and IQ. *Journal of Biosocial Science, 28*, 559-572.

Miao, E. S. Y. (1993). *Translation of J. Raven, J. C. Raven, & J. H. Court, Manual for Raven's Progressive Matrices Tests*. Taiwanese Edition. Taiwan: Chinese Behavioural Science Corporation.

Murphy, K. R., & Davidshofer, C. O. (1998). *Psychological testing*: *principles and applications* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Raven, J. (1981). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.1*: *The 1979 British Standardisation of the Standard Progressive Matrices and Mill Hill Vocabulary Scales, Together With Comparative Data From Earlier Studies in the UK, US, Canada, Germany and Ireland*. San Antonio, TX: Harcourt Assessment.

Raven, J. (2000a). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.3 (Second Edition): A Compendium of International and North American Normative and Validity Studies Together with a Review of the Use of the RPM in Neuropsychological Assessment by Court, Drebing, & Hughes*. San Antonio, TX: Harcourt Assessment.

Raven, J. (2000b). The Raven's Progressive Matrices: Change and stability over culture and time*. Cognitive Psychology, 41*, 1-48.

Raven, J., Raven, J. C., & Court, J. H. (1998, updated 2003) *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1*: *General Overview*. San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3*: *The Standard Progressive Matrices*. San Antonio, TX: Harcourt Assessment.

Raven, J. C. (1958). *The Standard Progressive Matrices.* London: H. K. Lewis. An earlier version of this test was known as *Progressive Matrices (1938)* and was also published by H. K. Lewis. The test was subsequently published by OPP Ltd. (Oxford) and now by Harcourt Assessment, San Antonio, TX.

Schaie, K. W., & Willis, S. L. (1986). *Adult Development and Ageing* (2nd edition). Boston: Little Brown.

Thorndike, R. L. (1975). *Mr. Binet's Test 70 Years Later*. Presidential Address to the American Educational Research Association.

Thorndike, R. L. (1977). Causation of Binet IQ decrements. *Journal of Educational Measurement, 14*, 197-202.

Chapter 14

The *Coloured Progressive Matrices* in South Africa

Adien Linstrom, John Raven, and Jean Raven
in collaboration with Jopie van Rooyen and Partners

Abstract

With a view to establishing adequate South African norms for Raven's *Coloured Progressive Matrices,* 2,469 children, aged 5 to 12 years and judged to form a representative sample of pupils of that age in the Free State were tested. As was the case in earlier, less broadly based, studies, the overall norms which resulted were somewhat lower than their UK equivalents. However, when the data were broken down by language of the home, it emerged that the norms for the English and Afrikaans speaking group were very similar to the UK norms. Those for the "other languages" group were higher than norms which have been reported for an Xhosa-medium primary school near Grahamstown.

*****

South Africa has eleven official languages and a population of approximately 45 million divided into four main groups. Black people represent a diversity of indigenous groups. White people are mainly descendents of European immigrants. Coloured people are descendents of cross-cultural relationships. And Asian people are mainly Indian. There are also minority groups of Chinese, Taiwanese, and Japanese. The population mix is unique in that Whites and Blacks have their roots in two totally different worlds. The first is primarily a European capitalistic industrialized society and the second mainly a pre-industrialized way of life.

Of the total population only an estimated 8% of adults have any post-school qualifications, 20% have school-leaving certificates, and 30% have some secondary school) (Statistics South Africa, 2003). It was with

this apparent need to find strategies to accelerate the education and development of the youth of South Africa in mind that the Education Department of the Free State asked Jopie van Rooyen and Partners to assist in establishing norms for the *Coloured Progressive Matrices*. (As explained in the General Introduction to this book, the Coloured Progressive Matrices (CPM) is designed to spread the scores of the bottom 20% of the population on the *Standard* Progressive Matrices. It consists of Set A and Set B of the Standard series, printed in color, with an additional set of items of intermediate difficulty. [See Raven, Raven and Court, 1998])

The Education department of the Free State province prepared the sample. This consisted of a random selection of schools from different regions, population groups, and school types chosen to yield as broad a database as possible. The research in the schools was carried out by Adien Linstrom in 2001. Data entry was organised by Jopie van Rooyen and Partners, but carried out by different institutions who did not always adhere to common instructions.

Completed CPM profiles were obtained from 2,469 South African children between the ages of 5 and 12 years. A breakdown of their socio-economic circumstances is given in Table 14.1.

One possible explanation of the apparent over-representation of children from rural areas might be that a greater proportion of children,

Table 14.1. Socio-Economic Composition of South African Sample of Young People Compared with State and National Statistics (Adults)

	Sample	Free State	South Africa
Gender			
Male	48%	48%	48%
Female	52%	52%	52%
Area			
Urban	62%	76%	58%
Rural	38%	24%	42%
Home Language			
Afrikaans/English	46%	13%	22%
Other	54%	87%	79%
Father's Status			
Professional	16%		
Technical	18%		
Administrative	14%		

in comparison with adults, live in rural areas – and this could well be the case, as rural families tend to be larger.

As far as the apparent over-representation of the Afrikaans and English speaking group is concerned, the vast majority of the population speak Indigenous African languages as a first language, and then English or Afrikaans as a second (or even third) language. However, it is becoming more common for parents to send their children to English or Afrikaans-medium schools, and the children often then speak English or Afrikaans at home as their first language.

Results

The overall South African norms derived from this sample are compared with the 1982 British Norms in Table 14.2.

When interpreting these data it should be borne in mind that there is much evidence to suggest that the British norms will have increased from 1982 to 2001.

Nevertheless the comparison is interesting. The figures for the 95th percentile for South Africa and the UK are similar while the scores for the lower percentiles drop increasingly behind.

In Table 14.3, the overall norms for the English and Afrikaan speaking group are compared with the overall norms for all other groups combined.

It will be seen that the norms for the English and Afrikaans speaking group are well above those for the combined "other languages" group and are, in fact, very similar to the 1982 British norms at all levels of ability. It may be worth commenting that similar results have been obtained when parallel analyses have been conducted within school districts in the USA.

Rather surprisingly, the norms for children living in rural areas do not differ much from those for children living in urban areas.

In common with the results obtained in many other societies, the norms for children coming from professional, administrative, and technical backgrounds are well above those for children whose fathers were labourers or who were unemployed.

Table 14.4 compares the Free State norms for both the (English plus Afrikaans) speaking group and the "All other languages" group with norms compiled from data collected by Natalie Bass from all children in

Table 14.2. *Coloured Progressive Matrices*
Smoothed 2001 Norms for South Africa In the Context of 1982 British Data

	Age in Years (Months)													
	5½ 5(3) to 5(8)		6 5(9) to 6(2)		6½ 6(3) To 6(8)		7 6(9) to 7(2)		7½ 7(3) to 7(8)		8 7(9) to 8(2)		8½ 8(3) to 8(8)		9 8(9) to 9(2)	
Percentile	UK	UK	**SA**	UK	**SA**	UK	**SA**	UK	**SA**	UK	**SA**	UK	**SA**	UK	**SA**
95	22	24	**23**	26	**25**	28	**27**	31	**29**	32	**30**	33	**31**	34	**32**
90	20	21	**19**	23	**21**	25	**24**	28	**26**	30	**28**	32	**30**	33	**31**
75	18	19	**16**	20	**17**	21	**19**	23	**21**	25	**23**	27	**25**	29	**27**
50	**15**	**16**	**12**	**17**	**13**	**18**	**14**	**20**	**15**	**22**	**16**	**24**	**17**	**26**	**19**
25	12	13	**9**	14	**10**	16	**11**	17	**11**	18	**12**	20	**13**	22	**14**
10	10	11	**7**	12	**8**	13	**8**	14	**9**	15	**9**	16	**9**	17	**10**
5	8	9	**6**	11	**6**	12	**7**	13	**7**	14	**7**	14	**7**	15	**7**
n	35	23	**56**	42	**108**	54	**232**	55	**220**	44	**186**	48	**226**	52	**211**

(continued)

	Age in Years (Months)											
	9½ 9(3) to 9(8)		10 9(9) to 10(2)		10½ 10(3) to 10(8)		11 10(9) to 11(2)		1½ 11(3) to 11(8)		12 11(9) to 12(2)	12½ 12(3) to 12(8)
Percentile	UK	**SA**	UK	**SA**	UK	**SA**	UK	**SA**	UK	**SA**	**SA**	**SA**
95	35	**33**	35	**33**	35	**34**	35	**34**	35	**35**	**35**	**35**
90	33	**32**	33	**32**	34	**33**	35	**33**	35	**34**	**34**	**34**
75	31	**28**	32	**29**	33	**30**	33	**31**	34	**32**	**33**	**33**
50	**28**	**21**	**30**	**23**	**31**	**25**	**31**	**26**	**32**	**27**	**29**	**30**
25	24	**14**	25	**15**	26	**16**	28	**17**	30	**20**	**22**	**25**
10	19	**11**	21	**11**	22	**12**	23	**13**	25	**14**	**15**	**17**
5	16	**8**	17	**9**	18	**10**	20	**11**	22	**12**	**13**	**16**
n	37	**212**	53	**191**	49	**218**	51	**190**	55	**216**	**105**	**94**

SA: South African data comprised of a sample of primary school children in the Orange Free State Province. The education department of the Free State randomly selected schools from different regions, population groups, and school types in an attempt to obtain a broad as possible data base.

UK: Based on a sample of 598 schoolchildren, including those attending special schools.

Table 14.3. *Coloured Progressive Matrices*
Smoothed 2001 Norms for South Africa By Language Spoken

	Age in Years (Months)														
	6		6½		7		7½		8		8½		9		9½	
	5(9) to 6(2)		6(3) to 6(8)		6(9) To 7(2)		7(3) to 7(8)		7(9) to 8(2)		8(3) to 8(8)		8(9) to 9(2)		9(3) to 9(8)	
Percentile	O	E+A	O	E+A	O	E+A	O	E+A	O	E+A	O	E+A	O	E+A	O	E+A
95	16	26	18	28	21	30	24	32	26	33	27	34	29	34	30	
90	15	24	16	26	17	27	19	29	21	30	23	32	25	33	27	
75	13	21	14	23	15	25	16	27	17	28	18	30	19	31	21	
50	**11**	**17**	**12**	**18**	**12**	**19**	**13**	**22**	**13**	**24**	**14**	**26**	**15**	**27**	**17**	
25	8	13	9	14	10	15	10	17	10	19	11	21	12	23	12	
10	7	10	7	11	8	12	8	13	8	14	9	15	9	16	10	
5	6	9	6	9	6	10	6	11	6	12	6	13	6	14	7	
n	36	61	60	96	122	94	121	75	116	100	119	90	109	83	120	

(continued)

	Age in Years (Months)											
	10		10½		11		11½		12		12½	
	9(9) to 10(2)		10(3) to 10(8)		10(9) to 11(2)		11(3) to 11(8)		11(9) to 12(2)		12(3) to 12(8)	
Percentile	E+A	O	E+A	O	E+A	O	E+A	O	E+A	O	E+A	O
95	34	31	35	32	35	32	35	33		34		34
90	33	28	34	30	34	31	34	32	35	33	35	34
75	32	23	32	25	33	27	33	28	34	30	34	32
50	**28**	**19**	**29**	**20**	**30**	**21**	**30**	**23**	**31**	**24**	**31**	**25**
25	24	13	25	13	26	16	27	17	28	18	29	20
10	18	10	20	11	21	11	22	12	23	14	24	15
5	15	7	16	8	17	9	18	10	19	12	19	13
n	71	111	87	123	75	113	90	119	51	52	36	59

E+A: English and Afrikaans speaking.

O: Other languages.

a Xhosa-speaking primary school in Joza, Grahamstown, South Africa, others compiled by Viki Costenbader in Kenya and the UK norms. It will be seen that the norms for the English-and-Afrikaans speaking group are similar to the UK norms, that the Joza Xhosa norms are lower than the Free State All-other-Languages group, and that Vicki Costenbader's Keynan norms are similar to the South African All-other-Languages group.

It has often been suggested that the difference between the norms for the Westernised groups and indigenous Africans (which are similar to those for indigenous Americans) might be, at least in part, due to the latter groups' relative unfamiliarity with the way of thought required to solve the problems presented in the test. To test this hypothesis, Nicola Taylor ran a 1-parameter IRT-based item analysis separately within these groups. The correlation between the item difficulties established separately in the English-speaking and Indigenous African group was .97. This is similar to the correlations obtained in other studies (see Jensen, 1998; Owen, 1992; Raven, 2000; and Raven et al., 2000, updated 2004) and seriously calls into question the hypothesis that the difference in mean score between the groups is due to one group's unfamiliarity with the way of thought required by the test. The test is working, and working in the same way, for both groups.

Table 14.4. *Coloured Progressive Matrices*
2001 Norms for the Orange Free State
Afrikaans + English Speaking and All-Other-Languages Groups' norms in the Context of 2001 Joza (Grahamstown) Xhosa-Speaking Group, Kenyan Norms and 1982 British Data

	Age in Years (Months)															
	5½	6			6½			7					7½			
	5(3) to 5(8)	5(9) to 6(2)		5(4) to 6(3)	6(3) to 6(8)			6(9) to 7(2)				6(4) to 7(3)	7(3) to 7(8)			
Percentile	UK	UK	**O**	**KN**	UK	**E+A**	**O**	UK	**E+A**	**O**	**XH**	**KN**	UK	**E+A**	**O**	**XH**
95	22	24	**16**	**17**	26	**26**	**18**	28	**28**	**21**	**16**	**20**	31	**30**	**24**	**17**
90	20	21	**15**	**16**	23	**24**	**16**	25	**26**	**17**	**15**	**18**	28	**27**	**19**	**16**
75	18	19	**13**	**14**	20	**21**	**14**	21	**23**	**15**	**14**	**15**	23	**25**	**16**	**15**
50	**15**	**16**	**11**	**12**	**17**	**17**	**12**	**18**	**18**	**12**	**12**	**13**	**20**	**19**	**13**	**13**
25	12	13	**8**	**10**	14	**13**	**9**	16	**14**	**10**	**11**	**11**	17	**15**	**10**	**12**
10	10	11	**7**	**7**	12	**10**	**7**	13	**11**	**8**	**10**	**9**	14	**12**	**8**	**11**
5	8	9	**6**	**6**	11	**9**	**6**	12	**9**	**6**	**9**	**8**	13	**10**	**6**	**10**
n	35	23	**36**	**237**	42	**61**	**60**	54	**96**	**122**	**4**	**213**	55	**94**	**121**	**6**

Table 14.4. *Coloured Progressive Matrices*
2001 Norms for the Orange Free State
Afrikaans + English Speaking and All-Other-Languages Groups' norms in the Context of 2001 Joza (Grahamstown) Xhosa-Speaking Group, Kenyan Norms and 1982 British Data *(continued)*

	Age in Years (Months)														
	8					8½					9				
	7(9) to 8(2)				7(4) to 8(3)	8(3) to 8(8)					8(9) To 9(2)				8(4) to 9(3)
Percentile	UK	**E+A**	**O**	**XH**	**KN**	UK	**E+A**	**O**	**XH**		UK	**E+A**	**O**	**XH**	**KN**
95	32	**32**	**26**	**19**	**25**	33	**33**	**27**	**20**		34	**34**	**29**	**21**	**30**
90	30	**29**	**21**	**18**	**23**	32	**30**	**23**	**19**		33	**32**	**25**	**20**	**28**
75	25	**27**	**17**	**16**	**18**	27	**28**	**18**	**16**		29	**30**	**19**	**17**	**22**
50	**22**	**22**	**13**	**14**	**14**	**24**	**24**	**14**	**14**		**26**	**26**	**15**	**15**	**15**
25	18	**17**	**10**	**12**	**11**	20	**19**	**11**	**13**		22	**21**	**12**	**13**	**12**
10	15	**13**	**8**	**11**	**9**	16	**14**	**9**	**12**		17	**15**	**9**	**12**	**10**
5	14	**11**	**6**	**10**	**8**	14	**12**	**6**	**11**		15	**13**	**6**	**11**	**9**
n	44	**75**	**116**	14	**255**	48	**100**	**119**	**19**		52	**90**	**109**	**27**	**289**

(continued)

	Age in Years (Months)													
	9½				10					10½				
	9(3) to 9(8)				9(9) to 10(2)				9(4) to 10(3)	10(3) To 10(8)				
Percentile	UK	**E+A**	**O**	**XH**	UK	**E+A**	**O**	**XH**	**KN**	UK	**E+A**	**O**	**XH**
95	35	**34**	**30**	**22**	35	**34**	**31**	**24**	**33**	35	**35**	**32**	**26**
90	33	**33**	**27**	**21**	33	**33**	**28**	**23**	**31**	34	**34**	**30**	**25**
75	31	**31**	**21**	**18**	32	**32**	**23**	**19**	**26**	33	**32**	**25**	**21**
50	**28**	**27**	**17**	**15**	**30**	**28**	**19**	**16**	**18**	**31**	**29**	**20**	**18**
25	24	**23**	**12**	**13**	25	**24**	**13**	**14**	**13**	26	**25**	**13**	**14**
10	19	**16**	**10**	**12**	21	**18**	**10**	**13**	**10**	22	**20**	**11**	**13**
5	16	**14**	**7**	**11**	17	**15**	**7**	**12**	**9**	18	**16**	**8**	**12**
n	37	**83**	**120**	**16**	53	**71**	**111**	**27**	**234**	49	**87**	**123**	**31**

E+A: English and Afrikaan speakers in South Africa. See Table 14.2 for a description of the sample.

O: Other languages in South Africa. See Table 14.2 for a description of the sample.

KN: Kenyan data collected by Virginia Costenbader and Stephen Mbugua Ngari from 1,370 children in the primary schools of the Municipality of Nakuru, a region which is fairly typical of the overall population of Kenya. 50% of the children were from the Kikuyu tribe and 21% Luo. The data have been re-worked by the authors. See Raven et al (1998).

UK: Based on a sample of 598 schoolchildren, including those attending special schools. See Raven et al (1998)

XH: The study was conducted by Natalie Bass. The sample was drawn from a representative Xhosa-medium Public Primary School in Joza, a township on the outskirts of Grahamstown, South Africa. See Bass (2000) and Knoetze et al (2005).

References

Bass, N. (2000). The Raven's Coloured Progressive Matrices Test: A Pilot Study for the Establishment of Normative Data for Xhosa-speaking Primary School Pupils in the Grahamstown Region. M.A. Thesis, Department of Psychology, Rhodes University, Grahamstown, South Africa.

Jensen, A. R. (1998). *The **g** Factor: The Science of Mental Ability.* Westport, CN: Praeger.

Knoetze, J., Bass, N. & Steele, G. (2005). The Raven Coloured Progressive Matrices: Pilot norms for isiXsa-speaking primary school learners in peri-urban Eastern Cape. *South African Journal for Psychology, 35*, 175-194.

Owen, K. (1992). The suitability of Raven's Standard Progressive Matrices for various groups in South Africa. *Personality and Individual Differences, 13*, 149-159.

Raven, J. (2000). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.3 (Second Edition): A Compendium of International and North American Normative and Validity Studies Together with a Review of the Use of the RPM in Neuropsychological Assessment by Court, Drebing, & Hughes*. San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 2: The Coloured Progressive Matrices*. San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3: The Standard Progressive Matrices*. San Antonio, TX: Harcourt Assessment.

Statistics South Africa. (2003). *Census 2001: Census in Brief* (Report Number 03-02-03). Pretoria, South Africa: SSA.

Chapter 15

Raven's *Standard* and *Advanced Progressive Matrices* among Adults in South Africa

Nicola Taylor

Due to the multicultural nature of the South African population and the fact that the country boasts 11 official languages, Raven's Standard (SPM) and Advanced (APM) Progressive Matrices are often used in organisational contexts as measures of cognitive ability. The emergence of the Employment Equity Act (55 of 1998) created a hesitance in the commercial sector with regard to the use of psychological assessments, as the Act clearly stipulates that psychometric assessments may not be used unless they have proven reliability and validity, are not biased against any employee, and can be fairly applied to any employee or group. The non-verbal nature of the SPM and APM lends them to the assumption of fairness, as language ability is excluded from the measurement of cognitive ability. However, questions have arisen as to whether these assessments measure the same construct in different groups, and whether the test is biased against individuals classed as previously disadvantaged through the apartheid system.

The present study was undertaken in order to investigate whether Raven's Standard and Advanced Progressive Matrices function similarly for Black and White working adults in the South African context. Item response theory, as conceptualised by the Rasch model (Rasch, 1960) was used to investigate whether or not the tests measure cognitive ability in the two groups in a similar way, and whether or not there is any evidence for bias in either the SPM or APM.

Standard Progressive Matrices (Classic Form)

The first research with the SPM in South Africa was by Rimoldi in 1945. This provided SPM percentiles for children aged 7 – 14 for each gender (Rimoldi, 1945). Since then, the SPM has proved to work effectively and

reliably in the South African context. However, most of the studies have focused on schoolchildren (e.g., Crawford-Nutt, 1976; Owen, 1992), and very few norm tables are available for adult samples in South Africa. Also, only a handful of published studies on the functioning of the Classic SPM items across cultural groups are available in South Africa (e.g., De Bruin, De Bruin, Derckson, & Cilliers-Hartslief, 2005).

The population studied in the current project

The data used for the analyses to be reported in this chapter were extracted from Jopie van Rooyen & Partners' (JvR) Consulting Services database. They were accumulated during selection exercises carried out for four major clients between 2005 and 2007. The demographics for the Classic SPM study are shown in Table 15.1. It will be seen that the data were provided by 144 female and 199 male job applicants, of whom 46.9% were Black and 41.8% White. The average age was 33.8 years. For the purposes of the following analysis, only the Black and White applicant groups will be compared, as the Indian and Coloured groups were too small.

Internal consistency

Although, as has been shown in earlier chapters, it is not entirely appropriate to calculate conventional measures of internal consistency for Item Response Theory (IRT) - based tests, these were generated to meet users' expectations.

Table 15.1. *Classic Standard Progressive Matrices*
South African Adult study
Demographics of the population studied

	Group	N	%
Gender	Women	144	33.8
	Men	199	46.7
	Unspecified	83	19.5
Total		426	100.0
Ethnicity	Black	200	46.9
	White	178	41.8
	Indian	32	7.5
	Coloured	4	0.9
	Unspecified	12	2.8

Table 15.2. *Classic Standard Progressive Matrices*
South African Adult Study
Internal consistency

Set	Number of items	Cronbach's alpha
Set A	10	0.75
Set B	12	0.87
Set C	12	0.85
Set D	12	0.86
Set E	12	0.86
Total Test	60	0.96

Table 15.2 shows the Cronbach alpha coefficients for each of the five sets of the Classic SPM, as well as that for the test as a whole. The internal consistency of Set A is lower than the others. This is because it consists mainly of easy items, which most adults are likely to get right. Also, as part of the demonstration of the answering procedure required, respondents are given the correct answers to the first two items on Set A, so these were removed from the reliability analysis. Overall, the internal consistency reliability of the SPM is very good.

Descriptive statistics

Table 15.3 presents descriptive statistics for the SPM for the entire group. As can be seen by looking at the mean scores for each set, respondents tend to get progressively fewer items correct the further they progress through the test. Overall, respondents got 45 out of the 60 items comprising the Classic SPM correct.

The skewness statistic gives an indication of how easy the respondents found the test. A negative skew means that most of the respondents got fairly high scores on the test. Figure 15.1 shows that the majority of scores on the SPM were at the higher end of the distribution. The kurtosis statistic is an indication of how sharp the curve of the distribution line is, or how concentrated the scores are. Positive kurtosis indicates a sharper spike in the curve than one would expect from a Gaussian – often misleadingly termed a "normal" – distribution. Figure 15.1 again illustrates that the scores are fairly highly concentrated around the mean score.

Group comparisons

Scores on the SPM were compared across gender and ethnic groups using an independent samples t-test. The results of the t-test for the gender

groups are presented in Table 15.4, and the results of the t-test for the ethnic groups are presented in Table 15.5.

The results of the t-test across gender groups show that there are no significant differences on any of the Sets of the Classic SPM, or on the total score.

Table 15.5 shows that, on all Sets, and on total score, the White group on average scored significantly higher than the Black group. This

Table 15.3. *Classic Standard Progressive Matrices*
South African Adult Study
Overall Descriptive Statistics by Set

Set	N	Min	Max	Mean	SD	Skewness	Kurtosis
Set A	426	4	12	10.95	1.59	-2.08	4.31
Set B	426	0	12	10.01	2.74	-1.84	2.60
Set C	426	0	12	8.80	2.93	-1.26	1.14
Set D	426	0	12	9.17	2.89	-1.75	2.64
Set E	426	0	12	5.73	3.40	.06	-1.12
Total score	426	6	60	44.65	11.94	-1.35	1.40

Figure 15.1. *Classic Standard Progressive Matrices*
South African Adult Study
Histogram of Total Scores

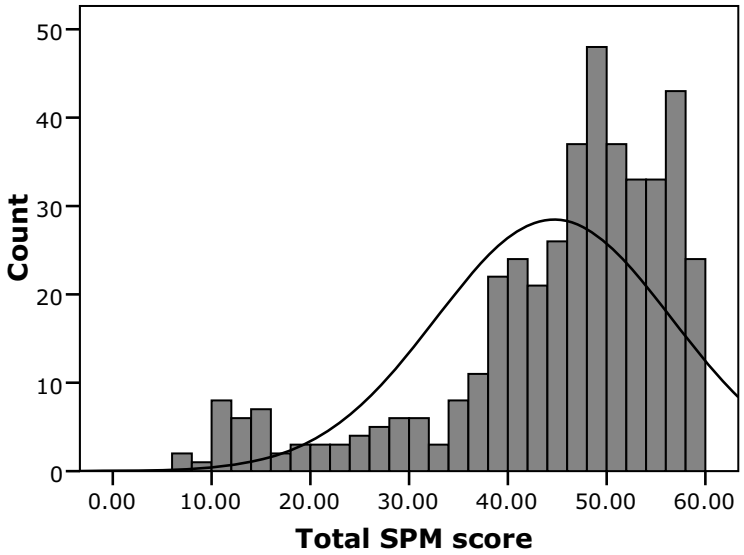


Table 15.4. *Classic Standard Progressive Matrices*
South African Adult Study
Mean Scores by Gender

	Men (N = 199)		Women (N = 144)			
Set	**Mean**	**SD**	**Mean**	**SD**	**t**	**P**
Set A	11.01	1.58	10.97	1.48	0.195	.846
Set B	9.90	2.92	10.07	2.55	-0.544	.587
Set C	8.87	3.05	8.60	2.92	0.845	.399
Set D	8.95	3.08	9.35	2.80	-1.209	.227
Set E	5.95	3.67	5.47	3.05	1.307	.192
Total score	44.69	12.64	44.45	11.28	0.183	.855

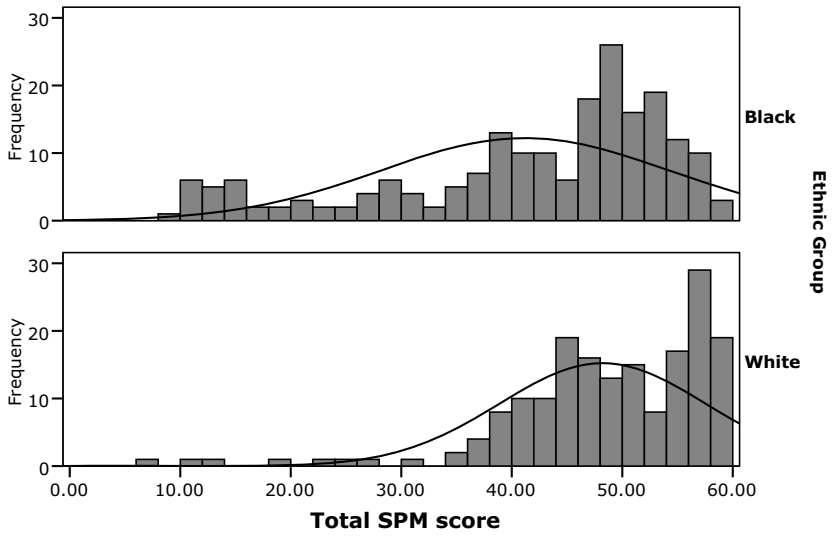
Table 15.5. *Classic Standard Progressive Matrices*
South African Adult Study
Mean scores by ethnicity

	Black (N = 200)		White (N = 178)			
Set	**Mean**	**SD**	**Mean**	**SD**	**t**	**P**
Set A	10.61	1.86	11.37	.98	-4.881	.000
Set B	9.45	3.17	10.62	1.98	-4.232	.000
Set C	8.12	3.09	9.50	2.60	-4.664	.000
Set D	8.43	3.28	9.89	2.19	-5.041	.000
Set E	4.60	3.16	6.84	3.33	-6.714	.000
Total score	41.20	13.06	48.21	9.33	-5.940	.000

finding may cause some concern at first, but it is important to consider the context in which the test was administered.

Figure 15.2 shows the separate score distributions for the Black and White applicant groups. From the graph, it can be seen that there was more of a spread of scores for Black applicants than for White applicants in terms of the total SPM score. This may well be due to the pre-selection process used to screen applicants before they reach the assessment phase. Most companies in South Africa are governed by Employment Equity policies, as well as Industry standards, which often predetermine the demographic profile of employees at different levels within the organisation. The shapes of the distribution of SPM scores (as seen in Figure 15.2) for Black and White applicants show that the distribution of the Black applicant group is more Gaussian than that for the Whites. This may be an indication that, on average, a select group of White

Figure 15.2. *Classic Standard Progressive Matrices*
South African Adult Study
Distribution of scores for Black and White applicants



applicants of higher ability are invited to participate in the assessment process, whereas the Black applicants who are invited to participate in the assessment process vary more widely in their level of ability as measured by the Classic SPM.

Differential item functioning

It has been widely argued that many psychological tests are "biased" against certain, often minority, groups. The meaning and interpretation of this claim has been highly contentious, often involving legal proceedings*. Despite repeated demonstrations (from 1938 onwards) that the RPM items scale in much the same way in most cultural groups (see Raven, 2000, for a review), the charge that it is unfair to certain groups continues to be levelled against it. The basis for this claim is that certain cultural groups "think in different ways" or, are, at the very least, "unfamiliar with the way of thought" required to perform well on the test. Yet, if this were the case, the items would not "scale"; their difficulty indices would be random, or, at the very least, when arranged in order of difficulty, the items would not be in the same order.

* Jensen (1980) offers a fairly thorough discussion of the issues.

Although this charge has repeatedly been shown to be unfounded, a number of tenable arguments and studies are still put forward to support this position. Most of these claims are based on attempts to apply the inappropriate assumptions of Classical Test Theory to a test which, as we have seen in other chapters, was built on, and conforms to, the requirements of, Item Response Theory. Nevertheless, because of the social significance of the claim in Africa, a serious effort was made to address the question (using the latest available techniques) in the course of the present study.

In the IRT literature, the question has been tackled under the rubric of "differential item functioning". Unfortunately, disputes at various conferences indicate that the term is not always understood in the same way and that there are disagreements about the best way of assessing "it".

Our own "take" on the issues is that there are two major groups of researchers. One is concerned solely with the question of whether the test as a whole "scales" in the same way among different cultural groups once the item difficulties are established using IRT based procedures … and, in the process, identifying those items which do not. The other group is concerned with whether, in addition, the individual Item Characteristic Curves have the same shape for both groups.

It follows from the material presented in the earlier chapter by Raven, Prieler and Benesch that the latter question can only be investigated if the ICCs are generated using a 3pl model. Since the present investigation was conducted using a Rasch model, often viewed as a 1pl model, this was without the scope of the present study.

The aim of those investigating differential item functioning is to find out whether test takers who have similar knowledge (as determined from total test scores) perform in similar ways on individual test questions regardless of their gender, age, or ethnicity.

The main premise of item response theory is that the higher a person's ability is relative to the difficulty of an item, the higher the probability of a correct response on that item. The item response theory model used to analyse the data in the present sample was the Rasch model, using the WINSTEPS® program version 3.58.1 (Linacre, 2005). The Rasch model is the only measurement model that can transform human data into abstract, equal-interval scales, while maintaining strict objective criteria for the construction of a scale that is separate from the sample distribution (Bond & Fox, 2001). As observed by Linacre (1996), "failure of a data set to fit the Rasch model implies that the data do not support the construction of measures suitable for stable inference" (p. 512).

Using a Rasch analysis means that the estimated item difficulties are theoretically independent of the characteristics of the sample of persons taken from the population of interest. In many statistical approaches, knowledge of the sample distribution is required or assumed. In Rasch, the details of the sample distribution are generally unknown until after the analysis is completed. So, in most cases, the estimated item difficulties are statistically the same when whether the sample as a whole is high or low performing, central or dispersed, unimodal or multimodal, skewed or symmetrical. Of course, this ideal is never achieved to perfection, but it at least allows the researcher to make inferences about the test regardless of the distribution of the sample (J. M. Linacre, personal communication, August 2007).

In order to best illustrate this point, it is useful to examine the person-item map produced by a Rasch analysis. Because item difficulties and person abilities are calculated on the same scale (called "log-odds units" or "logits" in Rasch terminology), it is easy to see the sample characteristics compared to the item characteristics. Figure 15.3 shows the person-item map for the Classic SPM items and the respondents tested for the present study.

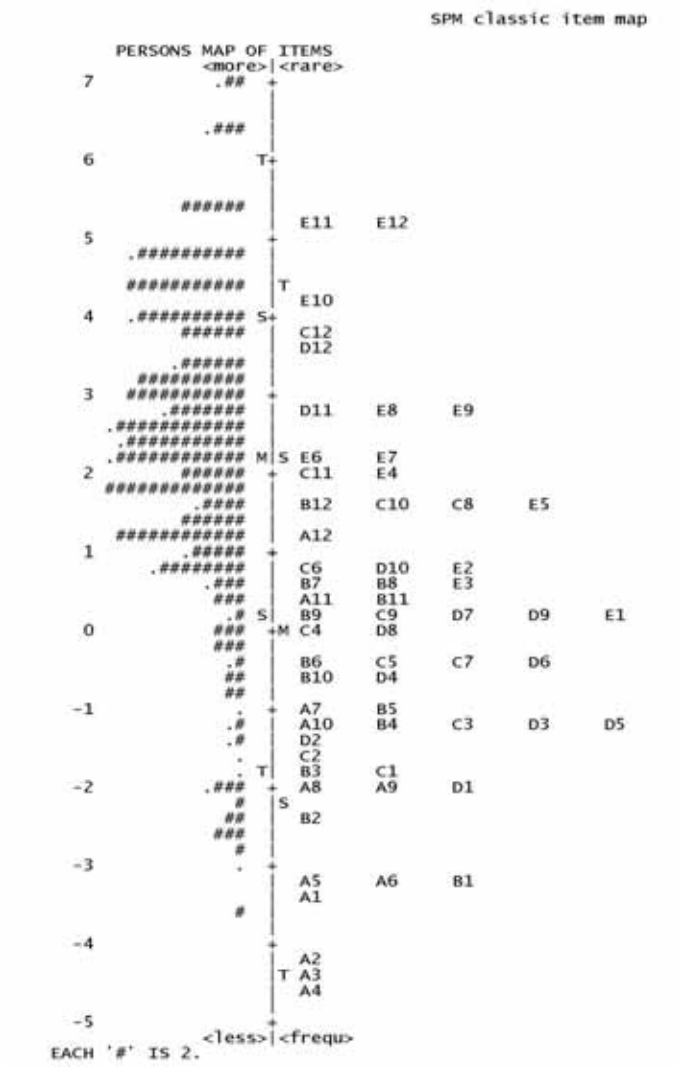
In Figure 15.3, the items appear on the right hand side of the line, distributed according to difficulty. The mean item difficulty for the items on the SPM is 0 logits – negative logits indicate easier items, and positive logits indicate more difficult items. The easiest items are A1, A2, A3, and A4, and the most difficult items are E11 and E12, as one would expect. The applicants appear on the left hand side of the map, distributed according to ability. The mean person ability is 2 logits, which is one standard deviation above the item difficulty. This is an indication that the items are actually too easy for many of the applicants, as can be seen by a number of applicants whose ability lies far above the highest level of ability tapped by the most difficult items, E11 and E12*.

As we have seen, if a test conforms to the Rasch model, the estimated item difficulties should be independent of the ability level of the sample which provided the basic data used to calculate them. Because the significance of this statement will be lost on many readers, attention may be drawn to just how starkly it differs from the situation that prevails when traditional indices of difficulty are employed. These show the proportion of the population tested who choose the correct answer to a given question.

* This, of course, is the reason why first the APM, and, later, the SPM ***Plus*** test discussed in other chapters, were developed.

Clearly, the proportion of a low ability sample which gets an item right is very much lower than the proportion of a high ability sample which gets it right. While these item difficulties may rank-order the items in the same way, the difficulty indices themselves are anything but identical.

Figure 15.3. *Classic Standard Progressive Matrices*
South African Adult Study
Person-item map



When using the Rasch model to seek evidence of differential item functioning between two groups, the question is, therefore, "To what extent is it true that the actual item difficulties, expressed in logits, are different for the samples which yielded the data on which they are based?"

The Rasch item difficulties assessed separately for the Black and White groups have therefore been plotted side by side in Figure 15.4 with a view to identifying those items which may not be functioning in the same way in both groups. In principle, these can be identified using a t-test … although, by definition, since there are 60 items in the SPM a proportion will have significantly different difficulty indices for purely statistical reasons.

It is clear that, in general, the items are functioning in very similar ways: Only four items could be flagged as having statistically significant difficulty indices (i.e. items B7, B8, C4 [which the White respondents found relatively more difficult], and E5 [which the Black respondents found relatively more difficult]). More detailed analyses would be required to determine whether these statistically significant differences reflect meaningful differences in psychological functioning.

The overall correlation between the Rasch-based item difficulty indices established separately in the Black and White groups was 0.97, again indicating that, despite the overall difference in average scores between the two groups, the test is functioning in an almost identical way within the two groups. It follows that common-sense-based "explanations" of the difference in mean score between the groups along the lines that "the test is unfamiliar to the way of thought of the Black group" do not hold up. The test works, and works in the same way, in both groups. Explanations of the difference must be sought elsewhere.

The overall functioning of the test

By now, we have established that the SPM is functioning in a similar way at the item level in both groups. It remains to ask whether the test as a whole functions in the same way for both groups.

We have encountered this question in our earlier chapter on *Lessons Learnt Whilst Developing a Romanian version of the MHV.* There, it was shown that one could in no sense assume that the overall Test Characteristic and Test Information Function curves developed in different settings would be similar, still less that these curves would, or should, conform to a Gaussian ogive*.

* Some readers may find it useful to be reminded that the Test Characteristic Curve is generated by summing the information contained in the individual Item Characteristic Curves.

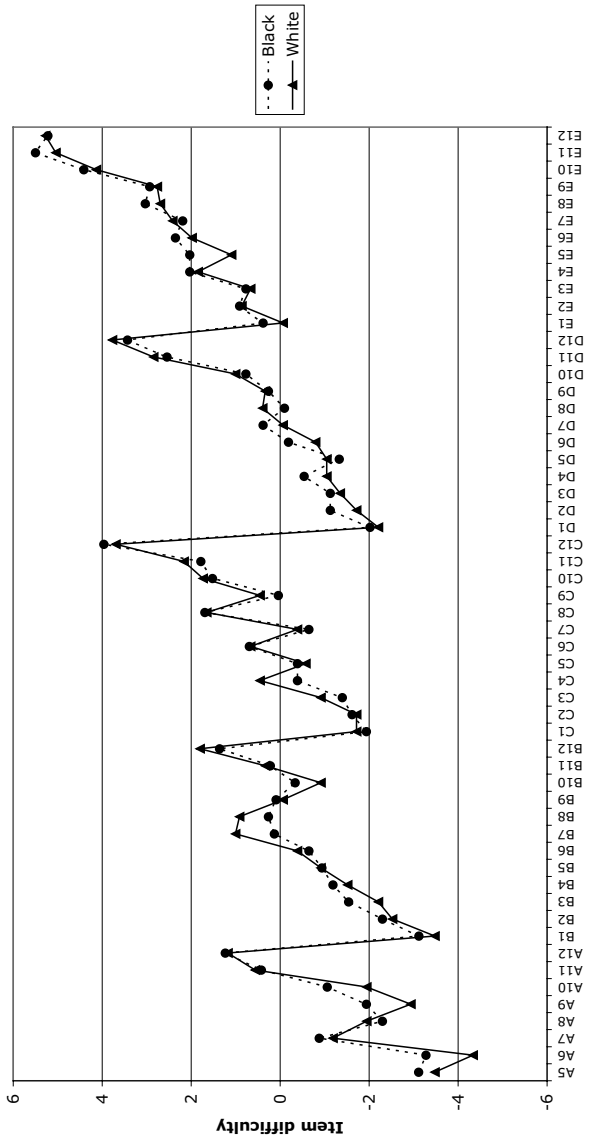


Figure 15.4. *Classic Standard Progressive Matrices*
South African Adult Study
Plot of item difficulties scaled separately for Black and White respondents.

In the light of these earlier discussions, the similarity between the Test Characteristic Curves calculated separately for the Black and White groups shown in Figure 15.5 is striking indeed.

Perhaps the most powerful statement that can be made on the basis of these results is that it would appear that Blacks and Whites who have the same ability as determined by their scores on the latent continuum have the same raw scores. The test is *not* "biased against them" because of their ethnicity.

Summary

The results of the analysis of the SPM data show that Black and White applicants of the same ability are equally likely to achieve the same score on the Raven's Standard Progressive Matrices in a selection context. The score differences obtained using Classical Test Theory methods are not a result of the differential functioning of the test, but more likely to be due to the composition of the sample. The SPM is a reliable measure of eductive ability, and the results of the present study are consistent with previous research.

The Advanced Progressive Matrices

Raven's Advanced Progressive Matrices (APM) was developed to differentiate between people of superior intellectual ability (Raven, Raven & Court, 1998). It is generally used in selecting employees for high-level technical or managerial positions. Research done on the APM in South Africa is limited, although some studies have been conducted in the South African National Defence Force (Muller & Schepers, 2003).

The APM has two components: Set I consists of 12 items that are often used as practice, as they provide training in the method of thinking required to complete Set II effectively. Set II consists of 36 items of increasing difficulty and constitutes the main part of the test (Raven, et al., 1998).

Populations Studied in the Present Project

The demographics for the APM group are shown in Table 15.6. 32 women and 158 men were involved, of whom 35.1% were Black and 60.2% White. The average age was 37.1 years. For the purposes of the following analysis, only the Black and White groups will be compared.

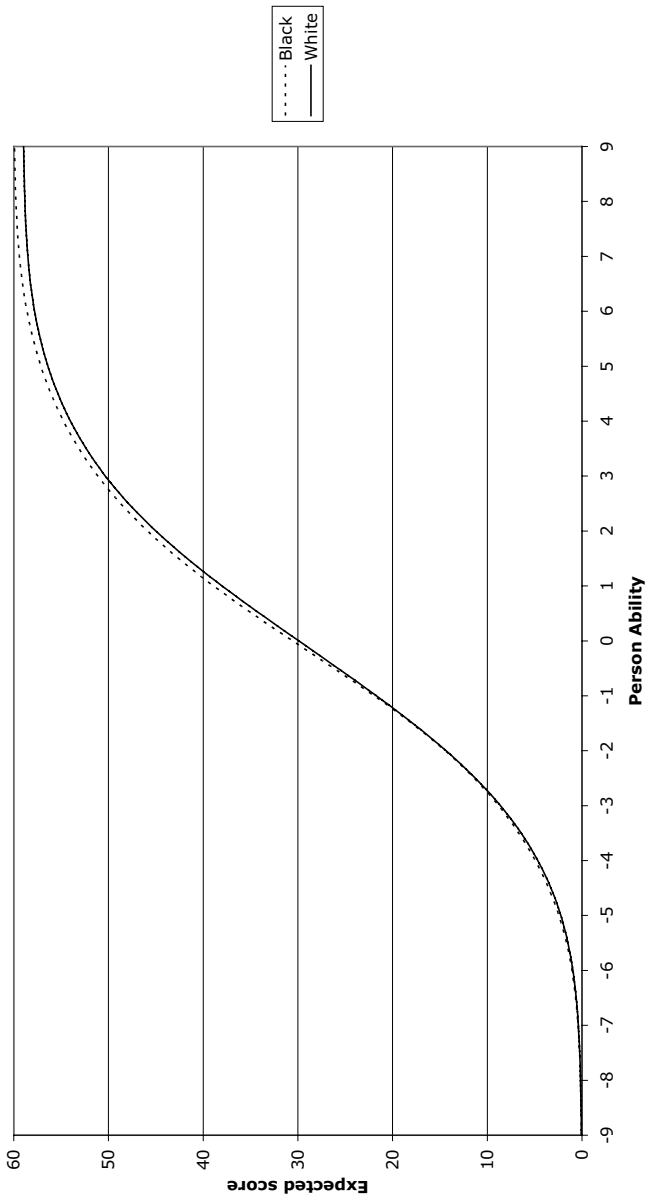


Figure 15.5. *Classic Standard Progressive Matrices*
South African Adult Study
Test characteristic curves for Black and White Respondents

Table 15.6. *Advanced Progressive Matrices.*
South African Adult Study
Demographics of Population Studied

	Group	N	%
Gender	Women	32	16.8
	Men	158	82.7
	Unspecified	1	0.5
Total		191	100.0
Ethnicity	Black	67	35.1
	Indian	8	4.2
	White	115	60.2
	Unspecified	1	0.5
Education	Grade 10	1	0.5
	Grade 12	11	5.8
	Higher diploma, National diploma, National certificate	80	41.9
	First degree, Honours degree	79	41.4
	Masters degree, Professional qualification	11	5.8
	Unspecified	9	4.7

Internal consistency reliability

The internal consistency reliability for Set I of the APM (12 items) is 0.57, which is low. The reason for this low reliability is probably due to the low difficulty level of the items in the first set, which is usually used as a training set, where candidates can clarify any of the items of which they are unsure. The internal consistency reliability for the Set II of the APM (36 items) is 0.89, which can be described as acceptable.

Descriptive statistics

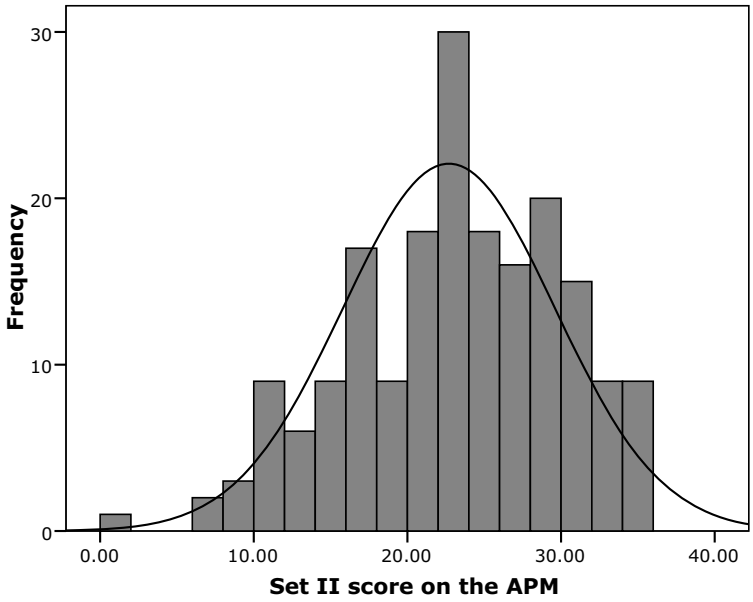
Table 15.7 presents the descriptive statistics on the APM for the entire group. As can be seen by looking at the mean scores for each set, respondents tend to get most of the Set I items correct. Overall, respondents got an average of 23 out of 36 items correct on Set II. From this point forward, only Set II scores will be included in analyses.

The skewness and kurtosis statistics for the Set II are very close to 0, indicating that the scores are distributed in a Gaussian curve. Figure 15.6 shows that the distribution of scores in graphic format.

Table 15.7. *Advanced Progressive Matrices.*
South African Adult Study
Descriptive statistics by Set

Set	N	Min	Max	Mean	SD	Skewness	Kurtosis
Set I	191	3	12	10.48	1.57	-1.73	4.81
Set II	191	1	36	22.70	6.90	-.36	-.31

Figure 15.6. *Advanced Progressive Matrices, Set II.*
South African Adult Study
Overall Score Distribution

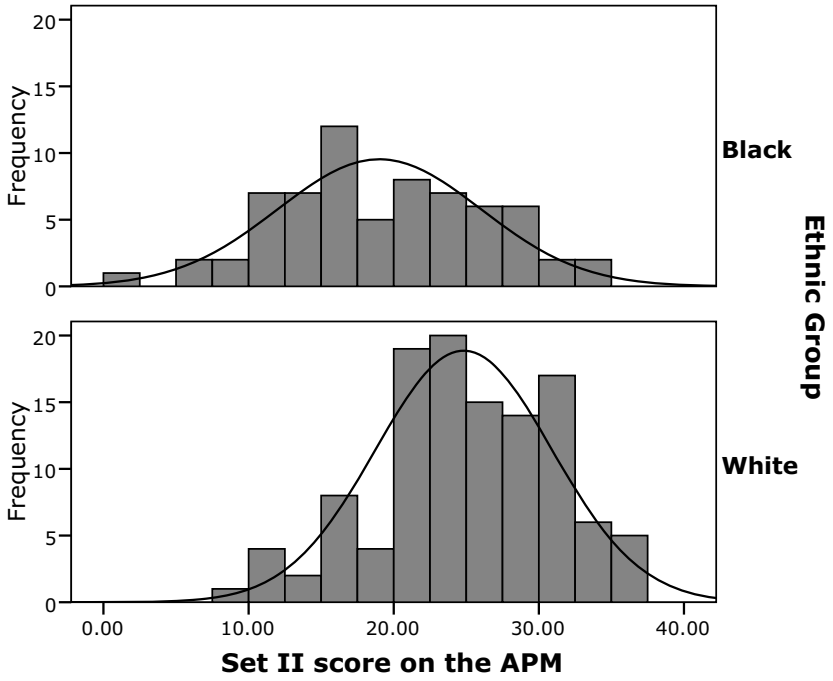


Group comparisons

Scores on the APM were compared across gender groups and ethnic groups using an independent samples t-test. The results showed no significant difference between men (Mean = 22.56, SD = 7.23) and women (Mean = 23.41, SD = 5.16) on the Set II APM [$t(188) = -0.632$, $p = 0.528$]. However, there was a significant difference between Black applicants (Mean = 19.01, SD = 7.01) and White applicants (Mean = 24.83, SD = 6.08) [$t(180) = -5.872$, $p = 0.000$].

The separate score distributions in Figure 15.7 show a similar pattern to the Classic SPM distributions across ethnic groups. Again, it is highly

Figure 15.7. *Advanced Progressive Matrices, Set II*
South African Adult Study
Score Distributions by Ethnicity



likely here that the difference in score distributions is due to the same selection practices employed in the use of the SPM.

Differential item functioning

In Figure 15.8, the items appear on the right hand side of the line, distributed according to difficulty. The mean item difficulty for the items on the APM Set II is 0 logits. The easiest items are 1, 2, and 3 and the most difficult items are 32 and 36, as one would expect. The applicants appear on the left hand side of the map, distributed according to ability. The mean person ability is just less than 1 logit, which is within one standard deviation above the mean item difficulty. This is an indication that, for the most part, the items are fairly well matched to the ability of applicants. All the applicants are more able than level tapped by the first three items, although there are some applicants whose ability lies above the level tapped by the most difficult item, number 36.

Figure 15.8. *Advanced Progressive Matrices, Set II*
South African Adult Study
Person-item map

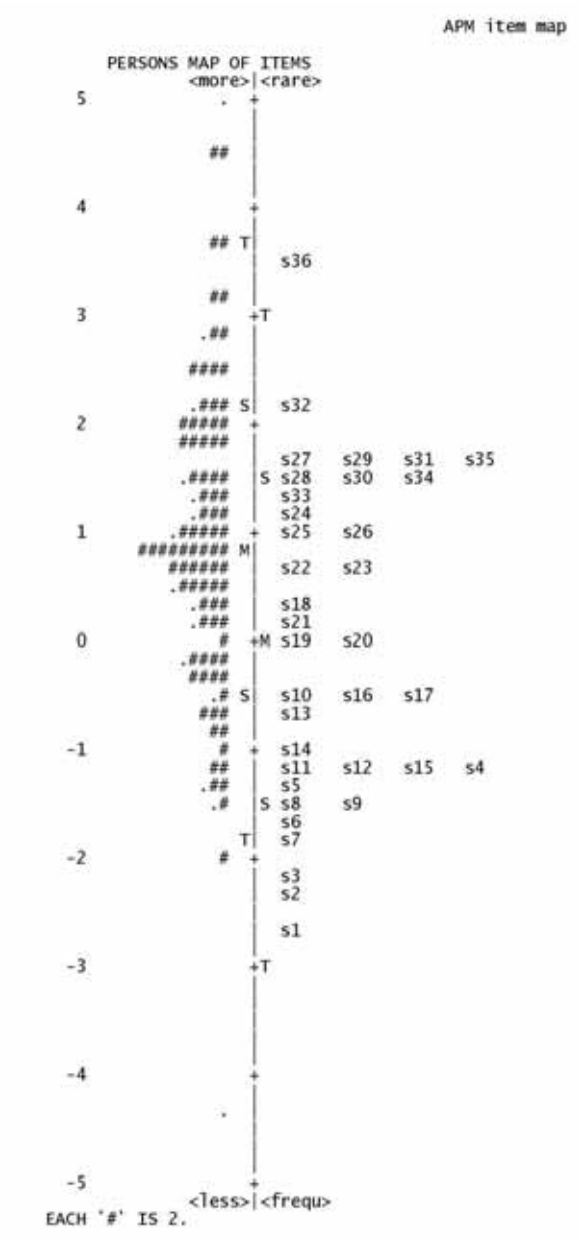


Figure 15.9 plots the Rasch item difficulties for Black and White applicants on Set II of the APM. The graph shows slightly different patterns of item difficulty for the two groups, but they show the same tendency that the items get more difficult as they progress through the test. Only five items could be flagged as possibly reflecting Differential Item Functioning. These were items 9 and 10, (which the Black applicants found more difficult), and 17, 19 and 28 (which the White applicants found more difficult). Although it appears that there may be differences in item difficulty for other items in Set II, they are not significant. The correlation between the item difficulties (in logits) determined separately among Black and White respondents was 0.93, which indicates that, despite the overall difference in average scores between the two groups, the test is functioning in an almost identical way within the two groups.

The Test Characteristic Curve (TCC) (calculated by cumulating the individual Item Characteristic Curves) gives an indication of what the expected raw score that someone having a given level of ability would be expected to attain. The TCC's for the Black and White applicants are shown in Figure 15.10. The resulting TCC's show that the curves are very similar, indicating that individuals of the same ability from either group are equally likely to obtain similar raw scores on the APM Set II.

Summary

The results of the analysis of the APM data show that Black and White applicants of the same ability are likely to achieve a similar score on the Raven's Advanced Progressive Matrices in a selection context. The raw score differences are again not a result of the differential functioning of the test, rather the composition of the sample.

Some conclusions

The results of the analysis of both the Classic SPM and APM data indicate that the claim that Black and White South Africans perform differently on these two tests is unsubstantiated. Regardless of cultural group, individuals of a certain ability level should be able to obtain the same raw score as others of the same ability.

The finding that the Black group did score lower on average than the White group is most likely a function of the sample characteristics. A larger, more representative sample is required before inferences are made as to why the nature of the samples differs.

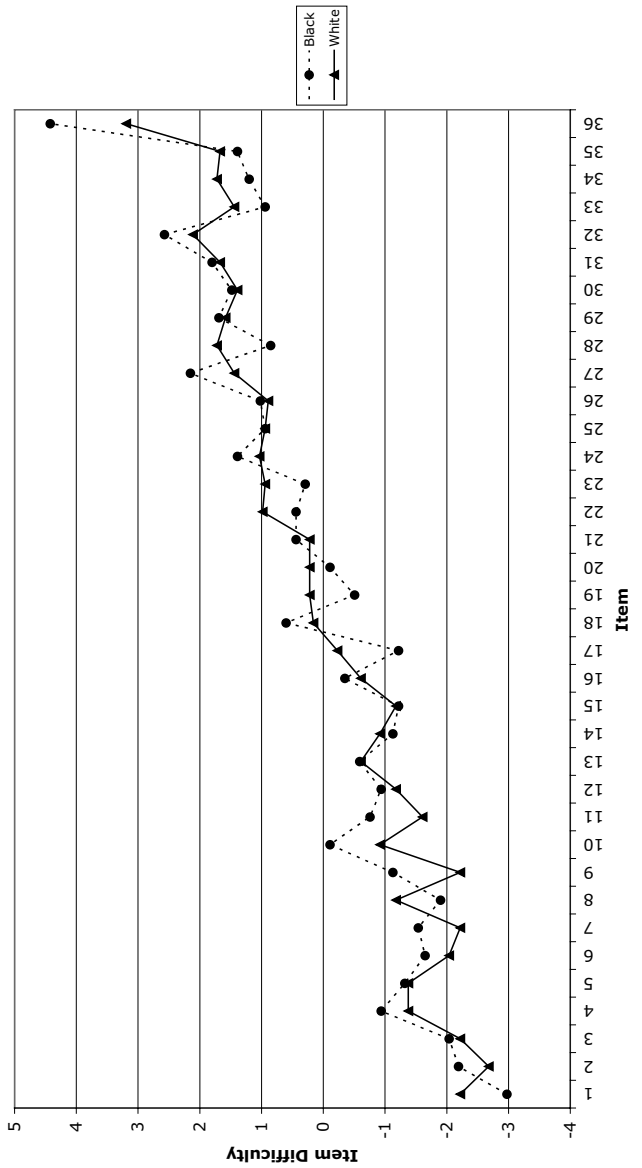


Figure 15.9 . *Advanced Progressive Matrices, Set II*
South African Adult Study
Plot of item difficulties calculated separately for Black and White respondents.

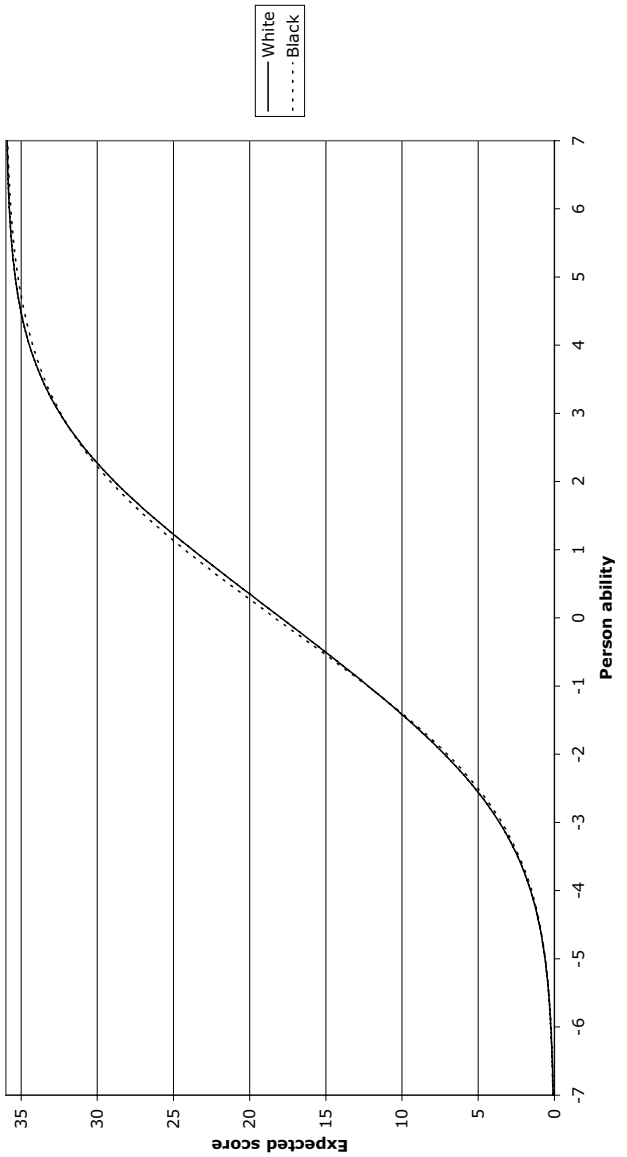


Figure 15.10. *Advanced Progressive Matrices, Set II*
South African Adult Study
Test characteristic curves for Black and White respondents.

References

Bond, T.G. & Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Crawford-Nutt, D.H. (1976). Are Black scores on Raven's Progressive Matrices an artifact of method of test presentation? *Psychologia Africana, 16,* 201-206.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-344.

De Bruin, K., De Bruin, G.P., Derckson, S., & Cilliers-Hartslief, M. (2005). Predictive validity of general intelligence and Big Five measures for adult basic education and training outcomes. *South African Journal of Psychology, 35,* 46-57.

Jensen, A. R. (1980). *Bias in Mental Testing*. New York: Free Press.

Linacre J.M. (1996). The Rasch Model cannot be "Disproved"! *Rasch Measurement Transactions,* 10, 512-514.

Linacre J.M. (2005). *WINSTEPS® Rasch Measurement.* Available from www.winstepscom.

Muller, J., & Schepers, J. (2003). The predictive validity of the selection battery used for junior leader training within the South African national defence force. *SA Journal of Industrial Psychology, 29,* 87-98.

Owen, K. (1992). The suitability of Raven's Standard Progressive Matrices for various groups in South Africa. *Personality and Individual Differences, 13* (2), 149-159.

Rasch, G. (1960). *Probalistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedagogiske Instituut.

Raven, J., Raven, J. C., & Court, J. H. (1998, updated 2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1: General Overview.* San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3: The Standard Progressive Matrices, Including the Parallel and Plus Versions.* San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 4: The Advanced Progressive Matrices.* San Antonio, TX: Harcourt Assessment.

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1-48.

Rimoldi, H.J.A. (1948). A note on Raven's progressive matrices test. *Educational and Psychological Measurement, 8,* 347-352.

Chapter 16

Standard Progressive Matrices Norms for Indian Tribal Areas

C. G. Deshpande and Vanita Patwardhan

This project was jointly conducted by
the Human Education Society, Pune
and
Jnana Prabodhini, Pune

This project was jointly funded by the British Research Foundation, U.K., and Manasayan, New Delhi.

A detailed report on the study is available from Manasayan, S-524 School Block, Shakarpur, Main Vikas Marg, Delhi – 110092.

Correspondence should be addressed to John Raven jraven@ednet.co.ukortoEmeritusProf.C.G.Deshpandeat<mangeshcd@rediffmail.com>

Acknowledgements

We gratefully acknowledge the help and assistance of: Dr. Usha Khire, Dr. P.H. Lodhi, Prof. D.J. Darekar, Dr. Arun Bacchav, Dr. Ashok Borse, Dr. V.S. Ainchwar, Dr. A.P. Shukla, Prof. R.D. Helode, Prof. B.A. Parikh, Prof. D.J. Darekar, Dr. Savita Deshpande, Dr. Sujala Watve, Mrs. Jyostna Joshi, Mrs. S. Nirgudkar, Dr. John Raven, Mrs. C.J. Raven, endless supervisors, data processors, and, of course all those who contributed the data on which the study was based.

Abstract

Data for the Standard Progressive Matrices were collected from representative samples of school attendees in nine tribal areas in India. Altogether, 6199 young people aged 8 to 21 were tested. The norms were considerably lower than both the Indian urban norms collected some years ago and those for the UK. Nevertheless, as usual, the most striking finding is the huge variance in ability within the groups: Poor educational and economic backgrounds do not have the extreme debilitating effect that many would have expected.

Introduction

India is a vast country with a plurality of cultures and many mature languages. The difference between rural and urban cultures has diminished over recent years. Nevertheless, hundreds of tribal areas still exist in many parts of the country, either far remote from the cities or isolated by geographical barriers such as forests. The means of communication are traditional and there is meagre economic development (most live in huts made of mud which are washed away in the rainy seasons), limited ways of earning a livelihood, and lack of education.

Since the cultural, educational, and economic differences are obvious, it is of considerable interest to know how the *Ravens Progressive Matrices* scores of young people living in these areas compare with those of their urban and international counterparts.

The Sample

Data were collected by trained researchers following the procedures laid down in the Manual from a sample of 6,199 school pupils. This was made up of between 650 and 700 students studying in grades 4-12 in each of nine tribal clusters. Altogether, data were collected in 72 schools. While there are hundreds of tribal areas all over the country, accessibility and funding meant that it was necessary to restrict the scope of the study to nine clusters: Thane, Pune, Raigadh, Solapur, Nashik, Dhule, Chandrapur, Surat and Raipur. Testing was carried out between November 2005 and March 2006.

Not all young people attend school, and the participation rates for females were much lower than those for males, especially in the higher grades. To the extent that young people not attending school were not included in the study, the results are less informative than might have been the case.

Most of the pupils' parents had not attended school at all.

Results

The overall smoothed summary norms from the study are presented in Table 16.1 and, broken down by sex in Table 16.2.

As can be seen from Table 16.2, there was a marked gender difference in favour of males which is probably due to differential participation rates.

As can be seen from Tables 16.3 and 16.4, the tribal norms were well below both the Indian urban norms and the 1979 UK norms (which are similar to the Indian urban norms).

To, yet again, investigate whether the differences in scores between the groups might be explained by failure of the test to work, item difficulties were calculated separately for the areas which were socio-economically closest to and furthest from Urban India and plotted in graph form. Both graphs corresponded closely to those obtained in other studies, such as the 1979 British standardisation with which the Indian Tribal norms have been compared in the above tables.

Table 16.1. *Standard Progressive Matrices*
Smoothed 2006 Norms for Indian Tribal Areas

	Age in Years							
Percentile	8½	9	9½	10	10½	11	11½	12
95	34	36	37	38	39	39	39	40
90	30	31	33	34	34	35	36	38
75	21	22	23	24	24	25	28	31
50	**15**	**15**	**16**	**16**	**16**	**17**	**19**	**20**
25	12	12	12	12	12	12	13	13
10	9	10	10	10	10	10	10	10
5	8	8	8	8	8	8	8	8
n	39	84	206	245	309	245	361	284

Percentile	12½	13	13½	14	14½	15	15½	16
95	41	43	44	45	46	47	48	49
90	39	40	41	42	43	44	45	46
75	33	34	35	36	37	37	39	40
50	**22**	**23**	**24**	**26**	**27**	**29**	**30**	**32**
25	13	13	14	15	15	17	19	20
10	10	10	10	11	11	12	12	12
5	8	8	9	9	9	10	10	11
n	426	320	463	287	449	341	479	262

Percentile	16½	17	17½	18	18½	19	19½
95	49	49	48	48	47	46	46
90	46	46	45	45	45	44	44
75	40	41	41	41	41	40	39
50	**33**	**34**	**35**	**35**	**35**	**34**	**33**
25	22	24	26	28	29	28	27
10	13	14	15	16	17	16	15
5	11	11	11	12	13	12	12
n	352	243	251	131	144	83	87

Note: Ages are those shown plus up to six months.

Table 16.2. *Standard Progressive Matrices* **Smoothed 2006 Norms for Indian Tribal Data – Female /Male Comparisons**

	Age in Years (Months)													
	9		9½		10		10½		11		11½		12	
Percentile	F	M	F	M	F	M	F	M	F	M	F	M	F	M
95	34	39	35	40	35	40	35	40	37	41	37	41	38	42
90	28	34	29	35	30	36	31	36	32	37	34	37	36	39
75	18	25	19	25	20	26	21	27	23	29	24	31	26	33
50	**13**	**17**	**14**	**17**	**15**	**18**	**15**	**18**	**15**	**18**	**16**	**20**	**17**	**22**
25	12	12	12	12	12	13	12	13	12	13	12	13	12	14
10	10	10	10	10	10	10	10	10	10	10	10	10	10	11
5	8	8	8	8	8	8	8	8	8	8	8	8	9	9
n	29	55	95	111	127	118	147	162	124	121	145	216	143	141

	13		13½		14		14½		15		15½		16	
Percentile	F	M	F	M	F	M	F	M	F	M	F	M	F	M
95	41	44	43	44	44	45	45	46	46	47	47	48	48	49
90	39	41	40	41	41	42	42	43	43	44	44	44	45	45
75	31	37	33	37	34	37	35	39	37	39	39	39	41	39
50	**20**	**24**	**21**	**25**	**22**	**27**	**24**	**29**	**26**	**30**	**28**	**32**	**30**	**32**
25	13	15	14	16	14	16	15	18	16	19	18	20	19	22
10	10	11	10	11	11	11	11	11	11	12	11	13	12	14
5	9	9	9	9	9	9	9	10	9	10	9	10	11	10
n	180	140	246	217	149	138	221	228	153	188	201	278	118	144

Note: Ages are those shown plus up to six months.

(continued)

Table 16.2. *Standard Progressive Matrices* **Smoothed 2006 Norms for Indian Tribal Data – Female /Male Comparisons** *(continued)*

	Age in Years (Months)													
	16½		17		17½		18		18½		19		19½	
Percentile	F	M	F	M	F	M	F	M	F	M	F	M	F	M
95	48	50	48	50	48	49	47	48	47	48	47	47	46	46
90	45	47	45	47	44	46	44	46	44	46	44	45	40	45
75	41	40	41	41	41	41	41	41	40	41	37	40	35	40
50	**32**	**34**	**33**	**36**	**34**	**36**	**35**	**36**	**35**	**37**	**35**	**34**	**32**	**34**
25	20	24	21	27	24	28	26	29	27	29	28	27	27	26
10	12	14	14	15	14	15	16	16	17	18	16	17	13	16
5	11	11	12	11	12	11	12	12	13	13	12	13	11	14
n	151	201	101	142	87	164	61	70	64	80	35	48	29	58

Note: Ages are those shown plus up to six months.

Table 16.3. *Standard Progressive Matrices*
Smoothed 2006 Norms for Indian Tribal Areas In the Context of 1997 Norms for Pune and Mumbai (Bombay), India.

	Age in Years (Months)										
	8	8½	9	9	9½	10	10	10½	11	11	11½
Percentile	P&M	**TR**	P&M	**TR**	**TR**	P&M	**TR**	**TR**	P&M	**TR**	**TR**
95	39	**34**	44	**36**	**37**	46	**38**	**39**	49	**39**	**39**
90	36	**30**	41	**31**	**33**	43	**34**	**34**	46	**35**	**36**
75	31	**21**	34	**22**	**23**	37	**24**	**24**	41	**25**	**28**
50	**19**	**15**	**21**	**15**	**16**	**28**	**16**	**16**	**33**	**17**	**19**
25	13	**12**	13	**12**	**12**	17	**12**	**12**	22	**12**	**13**
10	11	**9**	11	**10**	**10**	12	**10**	**10**	14	**10**	**10**
5	10	**8**	10	**8**	**8**	11	**8**	**8**	12	**8**	**8**
n	100	**39**	592	**84**	**206**	1104	**245**	**309**	1189	**245**	**361**

	12	12	12½	13	13	13½	14	14	14½	15	15
Percentile	P&M	**TR**	**TR**	P&M	**TR**	**TR**	P&M	**TR**	**TR**	P&M	**TR**
95	52	**40**	**41**	53	**43**	**44**	54	**45**	**46**	55	**47**
90	49	**38**	**39**	51	**40**	**41**	52	**42**	**43**	53	**44**
75	45	**31**	**33**	47	**34**	**35**	48	**36**	**37**	49	**37**
50	**39**	**20**	**22**	**41**	**23**	**24**	**43**	**26**	**27**	**44**	**29**
25	30	**13**	**13**	33	**13**	**14**	36	**15**	**15**	38	**17**
10	18	**10**	**10**	23	**10**	**10**	27	**11**	**11**	29	**12**
5	14	**8**	**8**	17	**8**	**9**	20	**9**	**9**	24	**10**
n	1293	**284**	**426**	1310	**320**	**463**	1344	**287**	**449**	1108	**341**

Note: TR ages are those shown plus up to six months.

(continued)

Table 16.3. *Standard Progressive Matrices*
Smoothed 2006 Norms for Indian Tribal Areas In the Context of 1997 Norms for Pune and Mumbai (Bombay), India.
(continued)

	Age in Years (Months)										
	15½	16	16	16½	17	17	17½	18	18	18½	19
Percentile	**TR**	P&M	**TR**	**TR**	P&M	**TR**	**TR**	P&M	**TR**	**TR**	**TR**
95	**48**	56	**49**	**49**	56	**49**	**48**	55	**48**	**47**	**46**
90	**45**	54	**46**	**46**	54	**46**	**45**	53	**45**	**45**	**44**
75	**39**	50	**40**	**40**	50	**41**	**41**	49	**41**	**41**	**40**
50	**30**	**45**	**32**	**33**	**45**	**34**	**35**	**44**	**35**	**35**	**34**
25	**19**	39	**20**	**22**	39	**24**	**26**	37	**28**	**29**	**28**
10	**12**	31	**12**	**13**	31	**14**	**15**	30	**16**	**17**	**16**
5	**10**	23	**11**	**11**	26	**11**	**11**	25	**12**	**13**	**12**
n	**479**	1192	**262**	**352**	769	**243**	**251**	287	**131**	**144**	**83**

Note: TR ages are those shown plus up to six months.

Table 16.4. *Standard Progressive Matrices*
Smoothed 2006 Norms for Indian Tribal Areas in the Context of 1979 British Data

	Age in Years										
	8	8½	8½	9	9	9½	9½	10	10	10½	10½
	7(9)	8(3)		8(9)		9(3)		9(9)		10(3)	
	to	to		to		to		to		to	
Percentile	8(2)	8(8)		9(2)		9(8)		10(2)		10(8)	
	UK	UK	**TR**	UK	**TR**	UK	**TR**	UK	**TR**	UK	**TR**
95	40	42	**34**	44	**36**	46	**37**	48	**38**	49	**39**
90	38	40	**30**	42	**31**	44	**33**	46	**34**	47	**34**
75	33	36	**21**	38	**22**	41	**23**	42	**24**	43	**24**
50	**25**	**31**	**15**	**33**	**15**	**36**	**16**	**38**	**16**	**39**	**16**
25	17	22	**12**	25	**12**	28	**12**	32	**12**	33	**12**
10	14	16	**9**	17	**10**	19	**10**	23	**10**	27	**10**
5	12	13	**8**	14	**8**	15	**8**	17	**8**	22	**8**
n	174	153	**39**	166	**84**	198	**206**	172	**245**	194	**309**

	Age in Years										
	11	11	11½	11½	12	12	12½	12.5	13	13	13½
	10(9)		11(3)		11(9)		12(3)		12(9)		13(3)
	to		to		to		to		To		to
Percentile	11(2)		11(8)		12(2)		12(8)		13(2)		13(8)
	UK	**TR**	UK	**TR**	UK	**TR**	UK	**TR**	UK	**TR**	UK
95	50	**39**	51	**39**	52	**40**	53	**41**	54	**43**	54
90	48	**35**	49	**36**	50	**38**	51	**39**	52	**40**	53
75	44	**25**	45	**28**	46	**31**	47	**33**	49	**34**	49
50	**40**	**17**	**41**	**19**	**41**	**20**	**42**	**22**	**43**	**23**	**44**
25	34	**12**	36	**13**	37	**13**	38	**13**	39	**13**	41
10	29	**10**	31	**10**	31	**10**	32	**10**	33	**10**	35
5	24	**8**	25	**8**	26	**8**	27	**8**	28	**8**	29
n	187	**245**	164	**361**	164	**284**	174	**426**	185	**320**	180

Note: TR ages are those shown plus up to six months. *(continued)*

Table 16.4. *Standard Progressive Matrices*
Smoothed 2006 Norms for Indian Tribal Areas in the Context of 1979 British Data *(continued)*

	Age in Years										
	13½	14	14	14½	14½	15	15	15½	15½	16	16½
		13(9)		14(3)		14(9)		15(3)			
		to		to		to		to			
Percentile		14(2)		14(8)		15(2)		15(8)			
	TR	UK	**TR**	UK	**TR**	UK	**TR**	UK	**TR**	**TR**	**TR**
95	**44**	55	**45**	56	**46**	57	**47**	57	**48**	**49**	**49**
90	**41**	54	**42**	54	**43**	55	**44**	55	**45**	**46**	**46**
75	**35**	50	**36**	50	**37**	51	**37**	51	**39**	**40**	**40**
50	**24**	**45**	**26**	**46**	**27**	**47**	**29**	**47**	**30**	**32**	**33**
25	**14**	42	**15**	42	**15**	42	**17**	42	**19**	**20**	**22**
10	**10**	36	**11**	36	**11**	36	**12**	36	**12**	**12**	**13**
5	**9**	30	**9**	33	**9**	33	**10**	33	**10**	**11**	**11**
n	**463**	196	**287**	189	**449**	191	**341**	171	**479**	**262**	**352**

	Age in Years					
Percentile	17	17½	18	18½	19	19½
	TR	**TR**	**TR**	**TR**	**TR**	**TR**
95	**49**	**48**	**48**	**47**	**46**	**46**
90	**46**	**45**	**45**	**45**	**44**	**44**
75	**41**	**41**	**41**	**41**	**40**	**39**
50	**34**	**35**	**35**	**35**	**34**	**33**
25	**24**	**26**	**28**	**29**	**28**	**27**
10	**14**	**15**	**16**	**17**	**16**	**15**
5	**11**	**11**	**12**	**13**	**12**	**12**
n	**243**	**251**	**131**	**144**	**83**	**87**

Note: TR ages are those shown plus up to six months.

Chapter 17

The *Standard Progressive Matrices* in Pakistan

Riaz Ahmad, Sarwat J. Khanam, and Zaeema Riaz

Abstract

The objective of the present study was to develop norms and conduct reliability and validity analyses for the Classic form of the Standard Progressive Matrices (SPM) in Pakistan. A sample of 1,662 Pakistani school students aged 11.11-18.11yrs responded to the test, which was administered, untimed, in group sessions. Pakistan is regionally divided into four provinces (NWFP, Baluchistan, Sindh & Punjab). In the present study, data were collected from the urban areas of each province. Besides the norms, the test's reliability and validity were estimated. The split half reliability was 0.89. In a separate validity study, SPM (Classic Form) scores were correlated with scores on the "Draw a Person" test for a sample of 200 school children aged 6 years 11 months to 11 years and 11 months. The correlation was 0.26. The Pakistani norms are compared with similar data accumulated in urban and tribal areas of India.

Introduction

The objective of the present study was to develop the norms, and to conduct reliability and validity analyses, for the Classic form of the *Standard Progressive Matrices* (SPM), which is widely used in Pakistan.

A particular incentive to conducting the study was Lynn's (1991) report that, while the average IQ (as assessed by General Intelligence Tests) of people living in Britain and the US is about 100, that of people living in North East Asia is around 105 and that of the peoples of Sub Saharan Africa around 70. In the light of such apparently large differences

between different nations it seemed unfair to compare an individual residing in Pakistan with norms developed in Britain. A standardization in Pakistan would make it possible to correctly assess the present level of functioning of individuals in the context of norms for the culture in which they survive.

Apart from the expected cultural differences between nations, even within Pakistan there are vast differences between the mental abilities of people from different areas. The environmental, economic and traditional differences between the people residing in different regions of Pakistan suggest a need for norms, taking into consideration the representation from all these areas.

Method

The Norm Sample

A sample (N=1,662) of Pakistani school students aged 11.11-18.11yrs responded to the Classic form of Raven's *Standard Progressive Matrices*. The test was administered, untimed, in group setting. The sample was stratified to ensure equal representation by age, sex, and race. However, some data were discarded for technical reasons during statistical analysis. An attempt was made to make the sample characteristics homogeneous.

Age

Six groups of adolescent children between the age ranges of 11 years 11 months to 18 years 11 months, with the mean age of 14 years 07 months, were selected. Table 17.2a indicates the age intervals and the number of subjects within each group by gender.

Regional Distribution

Pakistan is regionally divided into four provinces (NWFP, Baluchistan, Sindh & Punjab). In the present study data were collected from the urban areas of the capitals (Peshawar, Quetta, Karachi & Lahore) of each province. Other than the capitals, big cities from each province (except Baluchistan) including Rawalpindi and Islamabad (Punjab), Abbotabad (NWFP) and Hyderabad (Hyderabad), were also included in the sample to ensure the appropriate participation of all ethnic groups in the sample. Representation from Baluchistan is limited only to its capital due to

tribal influences in other areas of Baluchistan. Table 17.2b indicates the breakdown of each region by age and gender.

Gender

Male and female were sampled equally in proportion, however, some data was discarded for different technical reasons, and therefore participation of both male and female is approximately equal, 51.32% male and 48.67% female (Table 17.2a).

Table 17.1 a*. Demographic Data by Socio-Economic Status

	Lower SES	Middle SES	Higher SES
Average monthly income	14,000 and below	14,000 to 30,000	30,000 and above
Most common educational level of parents	Nil / Primary / Middle / Matric/ Skilled vocational	Intermediate/ Bachelor/ Master Degree	Bachelor/ Master Degree
Most common occupations	1. Clerical/Sales/ Service	1. Clerical/Sales/ Service	1. Professionals
	2. Drivers/ Peons/ Soldiers	2. Lecturers/ teachers	2. Business personnel
	3. Laborers	3. Doctors	3. Bureaucrats
Most common family structure	Extended / Joint	Joint / Nuclear	Nuclear

**Development of Norms and Application of Wide Range Achievement Test 3 in Pakistan - Trends in Adolescence*. Riaz Ahmad, Zaeema Riaz, & Sarwat Jahan Khanam. Institute of Clinical Psychology, University of Karachi (2005) pp. 8-9.

Table 17.1 b. Profile of Participating Students

School	SES	School Type	Gender (M & F)
A	L	G	**Approximately Equal**
B	L	NG	**-**
C	M	G	**-**
D	M	NG	**-**
E	H	NG	**-**

L = lower SES; H = higher SES; M = middle SES;
G = Government School; NG = Non-government School;
M = male; F = female

Socio-Economic Level

A Demographic Information Form was established to determine the socio-economic level of the subjects. Three major components were used to determine the socio-economic status are (a) Father's and Mother's level of education, (b) Father's and Mother's occupation, and (c) Family income. Other variables (such as residential area, family structure, number of siblings, earning members in the family) were also considered (Table 17.1).

Another important area i.e., the school system (government and private) that determines the Socio-Economic Status (SES) of children in Pakistan was also considered in the present study. In most of the Pakistani government schools, due to their lower fee structure, most of the pupils belong to the lower and lower middle SES groups. The participation of children in the present study of both school systems was approximately equal. Both non-government schools and government schools were selected within predetermined SES areas. The following profile was thus created.

Procedure

Permission from the authorities was initially taken after providing information regarding the present project. The administrator of schools was provided a letter for consent describing the research project and inviting participation, along with a sample of the SPM (Response Book & Answer sheet). The students were briefed on the nature of the research and were asked for their consent. After establishing rapport, the SPM

Table 17.2 a. *Standard Progressive Matrices in Pakistan*
Sample Composition by Age Group and Gender

GROUP	AGE	MALE	FEMALE	TOTAL
GROUP I	11.11 – 12.11	108	111	219
GROUP II	12.11 - 13.11	143	103	246
GROUP III	13.11 - 14.11	144	113	257
GROUP IV	14.11 - 15.11	142	116	258
GROUP V	15.11 - 16.11	103	135	238
GROUP VI	16.11 - 17.11	90	107	197
GROUP VII	17.11 - 18.11	123	124	247
TOTAL	11.11- 18.11	853 (51.32 %)	809 (48.67%)	1662

was administered by a group of competent and trained psychologists. In every class, testing was carried out by a psychologist and an assistant. The test was given without time limits. Standard instructions from SPM manual were given to subjects. The test was administered to randomly selected participants in groups of 20 in a classroom setting. Only those participants were included who willingly volunteered to participate in this project. The testing was carried out during the years 2004-2006.

Table 17.2 b. Age Group with Gender Breakdown of Four Provinces of Pakistan (N=1662)

AGE GROUP	AGE		Provinces							
			SINDH		NWFP		BALUCHISTAN		PUNJAB	
			MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE
GROUP I	11.11-12-11	N	22	28	25	25	33	23	28	35
		%	44	56	50	50	59	41	45	55
GROUP II	12.11 - 13.11	N	27	23	31	30	41	25	44	25
		%	54	46	49	51	62	38	64	36
GROUP III	13.11 - 14.11	N	35	35	43	31	19	17	47	30
		%	50	50	58	42	47	53	61	37
GROUP IV	14.11 - 15.11	N	48	40	33	29	25	23	36	24
		%	55	45	53	47	52	48	60	40
GROUP V	15.11 - 16.11	N	13	37	27	30	38	38	25	30
		%	26	74	47	53	50	50	54	46
GROUP VI	16.11 - 17.11	N	7	38	25	25	40	23	18	21
		%	16	84	50	50	63	37	46	54
GROUP VII	17.11 - 18.11	N	40	60	27	29	21	23	35	12
		%	40	60	48	52	48	52	74	26
TOTAL	11.11- 18.11	N	192	261	211	199	217	172	233	177

* % rounded off

Table 17.2 c. Percentages of Norm Sample within Region by Grades

PROVINCES	GRADES					
	SIX	SEVEN	EIGHT	MATRIC	INTERME-DIATE	GRADUATE
NWFP	19	48	46	171	126	0
PUNJAB	12	41	61	211	80	5
BALUCHISTAN	0	89	61	36	187	16
SIND	7	30	48	196	69	103
TOTAL	38	208	216	614	462	124

Table 17.3. Descriptive Characteristics of Data*

	Age						
	TWELVE	**THIRTEEN**	**FOURTEEN**	**FIFTEEN**	**SIXTEEN**	**SEVENTEEN**	**EIGHTEEN**
N	219	246	257	258	238	197	247
Mean	31.89	32.37	36.74	37.86	40.47	40.65	41.31
Median	34.00	34.50	39.00	39.00	41.00	42.00	43.00
Std. Dev	10.29	10.87	10.69	9.74	10.33	9.67	10.46
Skewness	-0.32	-0.36	-0.86	-0.74	-0.53	-1.27	-0.90
Kurtosis	-0.44	-0.84	0.38	0.77	-0.23	1.94	0.60

* Figures are rounded off upto two decimals

Table 17.4 a. *Standard Progressive Matrices (Classic Form)*
Self-Administered or Group Test Norms for Adolescents in Pakistan (Smoothed) (N=1662)

	AGE						
Percentiles	11.11 -12.11	12.11 - 13.11	13.11 - 14.11	14.11 - 15.11	15.11 - 16.11	16.11 - 17.11	17.11 - 18.11
95	49	50	53	54	57	57	59
90	45	47	49	51	53	54	56
75	40	41	43	49	50	51	51
50	**32**	**33**	**36**	**38**	**40**	**42**	**43**
25	24	25	28	31	35	35	35
10	17	19	20	23	26	27	28
05	13	14	15	17	19	21	21
N	219	246	257	258	238	197	247

Table 17.4 b. *Standard Progressive Matrices (Classic Form)*
Self Administered or Group Test, Smoothed Pakistan Norms for Adolescents by Gender (N=1662)

	AGE													
	11.11 - 12.11		**12.11 - 13.11**		**13.11 - 14.11**		**14.11 - 15.11**		**15.11 - 16.11**		**16.11 - 17.11**		**17.11 - 18.11**	
Percentiles	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**
95	50	49	50	50	52	53	54	54	56	56	57	57	58	58
90	43	43	44	44	46	46	48	48	50	50	51	51	52	52
75	38	37	39	39	41	41	43	43	46	45	47	46	47	48
50	**33**	**33**	**34**	**34**	**36**	**36**	**38**	**38**	**41**	**40**	**43**	**41**	**43**	**42**
25	27	25	28	27	30	29	33	32	36	34	38	36	38	37
10	19	17	20	19	23	21	26	25	29	27	32	29	32	30
05	13	11	14	12	15	14	18	17	21	20	23	22	23	24
N	108	111	143	103	144	113	142	116	103	135	90	107	123	124

Table 17.4 c. *Standard Progressive Matrices (Classic Form)*
Smoothed Norms for Adolescents in Pakistan
In the Context of 1997 Norms for Pune and Mumbai (Bombay), India and 2006 Norms for Indian Tribal Areas*

	Age											
	12	12	12½	12½	13	13	13½	13½	14	14	14½	14½
Percentiles	P&M	TR	TR	**PK**	P&M	TR	TR	**PK**	P&M	TR	TR	**PK**
95	52	40	41	**49**	53	43	44	**50**	54	45	56	**53**
90	49	38	39	**45**	51	40	41	**47**	52	42	43	**49**
75	45	31	33	**40**	47	34	35	**41**	48	36	37	**43**
50	**39**	**20**	**22**	**32**	**41**	**23**	**24**	**33**	**43**	**26**	**27**	**36**
25	30	13	13	**24**	33	13	14	**25**	36	15	15	**28**
10	18	10	10	**17**	23	10	10	**19**	27	11	11	**20**
05	14	8	8	**13**	17	8	9	**14**	20	9	9	**15**
N	1293	284	426	**219**	1310	320	463	**246**	1344	287	449	**257**

	Age											
	15	15	15½	15½	16	16	16½	16½	17	17	17½	17½
Percentiles	P&M	TR	TR	**PK**	P&M	TR	TR	**PK**	P&M	TR	TR	**PK**
95	55	47	48	**54**	56	49	49	**57**	56	49	48	**57**
90	53	44	45	**51**	54	46	46	**53**	54	46	45	**54**
75	49	37	39	**49**	50	40	40	**50**	50	41	41	**51**
50	**44**	**29**	**30**	**38**	**45**	**32**	**33**	**40**	**45**	**34**	**35**	**42**
25	38	17	19	**31**	39	20	22	**35**	39	24	26	**35**
10	29	12	12	**23**	31	12	13	**26**	31	14	15	**27**
05	24	10	10	**17**	23	11	11	**19**	26	11	11	**21**
N	1108	341	479	**258**	1192	262	352	**238**	769	243	251	**197**

	Age					
	18	18	18½	18½	19	19½
Percentiles	P&M	TR	TR	**PK**	TR	TR
95	55	48	47	**59**	46	46
90	53	45	45	**56**	44	44
75	49	41	41	**51**	40	39
50	**44**	**35**	**35**	**43**	**34**	**33**
25	37	28	29	**35**	28	27
10	30	16	17	**28**	16	15
05	25	12	13	**21**	12	12
N	287	131	144	**247**	83	87

*Norms for Pune & Mumbai, India (1997) [P&M] and for Indian tribal areas (2006) [TR] from Deshpande, C.G., & Patwardhan, V. (2006). *previous chapter*

Scoring and Statistical Analysis

The protocols were scored according to the standard method of scoring Raven's Standard Progressive Matrices. The analyses were carried out using standard statistical packages.

Reliability

Split Half Reliability

As a measure of internal consistency, the correlation between odd-item and even-item scores was computed by Pearson's method. The correct items were added up for half rows by balancing technique i.e., total of first, fourth, fifth, eighth, ninth and twelfth row were added, other remaining rows were added for the second half. Then the two resulted sums were used as two halves in the analysis. This procedure was acquired due to progressive difficulty level in the items and in the sets, due to which adding up of odd and even items was not supposed to be that much adequate.

Reliability-Comparison with Other Studies

Numerous researchers have reported on the split half reliability of the Classic SPM (see Raven, Court and Raven, 2000, updated 2004, and Court and Raven, 1995). The figure reported in Table 17.5 (0.89) is well within the range of those reported in other well-conducted studies.

Validity

The Classic form of the *Standard Progressive Matrices* is reported to have good psychometric characteristics (Murphy & Davidshofer, 1998; Kline, 2000). Therefore, it has gained widespread acceptance and is used in many countries all over the five continents (Irvine & Berry, 1988). A

Table 17.5. *Standard Progressive Matrices (Classic Form)*
Split Half Reliability in Pakistan (Ages 11.11-18.11)

	MEAN		STD. DEVIATION		R	SIG
N	ODD	EVEN				
1551	18.1870	18.6380	5.3253	5.7216	0.891	0.000

huge body of published research bears on its validity (Gregory, 1992). Validity is a useful tool to assess the fact that any Psychological measure in use assesses the abilities for which it claims or purports. The most frequently used way to evaluate validity is to relate it to other measures having the same purpose. Harris (1959), while using the SPM with 100 Kindergarten children selected to be representative of the US urban occupational distribution, found a correlation with the Goodenough, revised, of 0.22 (dealing with raw scores). While evaluating validity of SPM, as the measure of intelligence or say, spatial abilities and all for which it claims for, its relation to DAP (Goodenough, 1926) would be of more importance than with other tests, as Draw a person test measures children's ability to draw the figure of a man, children's handling of quantitative and spatial concepts.

In the course of the present project, the Classic SPM was administered to a sample of 200 school children whose ages ranged from 6 years 11 months to 11 years 11 months, with equal representation of both males and females. The sample was approached at three elementary schools; each belonging to the residing area of one of three socioeconomic classes i.e. low, middle, and high in order to make the equal representation of all SES possible. All the subjects were required to give demographic information (teachers were interviewed in case of young subjects). The pupils also completed the Draw A Person Test (Goodenough, 1926) according to standard procedure. Instructions were given in groups as well as individually for better understanding.

Table 17.6. Correlation between *Raven Standard Progressive Matrices* and *Draw A Person* (DAP) Test
Pakistani Elementary School Children (Age 6.11-11.11)

Measures	Mean	Std. Dev	Pearson's r	Sig
Draw A Person Test	33.0390	7.6131		
Standard Progressive Matrices	36.7013	9.0555	0.256	0.025

References

Court, J. H., & Raven, J. (1995). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 7*: *Research and References*: *Summaries of Normative, Reliability, and Validity Studies and References to All Sections*. San Antonio, TX: Harcourt Assessment.

Elley, W. B., & Macarthur, R. S. (1962). The *Standard Progressive Matrices* as a cultural-reduced measure of general intellectual ability. As reported in Court, J. H., & Raven, J. (1995). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 7*: *Research and References*: *Summaries of Normative, Reliability, and Validity Studies and References to All Sections*. San Antonio, TX: Harcourt Assessment.

Evans, L. (1966). A Comparative study of the Wechsler Intelligence Scale for Children (performance) and Raven's Progressive Matrices with deaf children. *Teacher of the Deaf, 64*, 76-82.

Ganguly, A. K. (1967). An experimental study of the variation in concept formation ability of young adults due to socioeconomic status. (As reported in Court, J. H., & Raven, J. (1995). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 7*: *Research and References*: *Summaries of Normative, Reliability, and Validity Studies and References to All Sections*. San Antonio, TX: Harcourt Assessment.)

Goodenough, F. L. (1926). *Measurement of Intelligence by Drawings*. New York: World Book Company.

Gregory, R. J. (1992). *Psychological Testing*: *History, Principles, and Applications*. Boston: Ally & Bacon.

Harris, D. B. (1959). A note on some ability correlates of the Raven Progressive Matrices (1947) in the Kindergarten. *Journal of Educational Psychology, 50*, 228-229.

Irvine, S. H., & Berry, J. W. (Eds.) (1988). *Human Abilities in Cultural Context*. Cambridge: Cambridge University Press.

Kaplan, R. M., & Saccuzzo, D. P. (1997). *Psychological Testing*: *Principles, Applications, and Issues* (4th ed.). Pacific Grove: Brooks/Cole. (As reported in Abdel-Khalek, A. M., & Raven, J. (2006). Normative data from the standardization of *Raven's Standard Progressive Matrices* in Kuwait in an international context. *Social Behaviour and Personality*: *An International Journal, 34(4).* Also: http://wpe.info/papers_table.html)

Kline, P. (2000). *Handbook of psychological testing* (2nd ed.). London: Routledge.

Lynn, R. (1991). Race differences in intelligence: a global perspective. *Mankind Quarterly, 31*, 255-294.

Lynn, R., Allik, J., Pullman, H., & Laidra, K. (2004). Sex differences on the Progressive Matrices among adolescents: Some data from Estonia. *Personality and Individual Differences, 36*, 1249-1255. (As reported in Abdel-Khalek, A., & Raven, J. (2005, September 2). Normative data from the standardization of Raven's Standard Progressive Matrices in Kuwait in an international context. WebPsychEmpiricist. Retrieved September 2, 2005, from http://wpe.info/papers_table.html)

MacKintosh, N. J. (1998). *IQ and Human Intelligence*. Oxford: Oxford University Press.

Murphy, K. R., & Davidshofer, C. O. (1998). *Psychological testing*: *principles and applications* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Raven, J. (2000). The Raven's *Progressive Matrices*: Change and Stability over Culture and Time. *Cognitive Psychology, 41*, 1-48.

Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3*: *The Standard Progressive Matrices,Including the Parallel and Plus Versions*. San Antonio, TX: Harcourt Assessment.

Sinha, M. (1977). Validity of the Progressive Matrices Test. As reported in Court, J. H., & Raven, J. (1995). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 7*: *Research and References*: *Summaries of Normative, Reliability, and Validity Studies and References to All Sections*. San Antonio, TX: Harcourt Assessment.

Sorokin, B. (1954). Standardization and analysis of Progressive Matrices Test by Penrose and Raven. As reported in Court, J. H., & Raven, J. (1995). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 7*: *Research and References*: *Summaries of Normative, Reliability, and Validity Studies and References to All Sections*. San Antonio, TX: Harcourt Assessment.

Spearman, C. (1927). *The Nature of 'Intelligence' and the Principles of Cognition* (2nd ed.). As reported in Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1-48.

PART IV

Outstanding Conceptual and Measurement Issues

So far, while we have indicated some of the problems that derive from the use of arbitrary metrics – which include most tests developed according to Classical Test Theory – and shown how these can be ameliorated by deploying appropriate forms of IRT, and while we have hinted at some of the problems which arise from the adoption of what might be called "arbitrary measures" in personal and programme evaluation, we have said little about problems which stem from the adoption of an inappropriate conceptual framework for thinking about, and assessing, individual differences.

Although these problems will be addressed at some length in the chapters in this Section, the nature of those problems can perhaps be indicated by asking "Where would biologists have got to if they had sought to summarise all the variation between animals in terms of 1, 2, or 16 "variables" analogous to e.g. "intelligence", "eductive and reproductive ability" or the 16PF (examples of such variables might be 'dogginess', or 'crabbiness', or 'aggressiveness'), the variance in environments in terms of, say, 10 variables (such as 'succorance' or 'animal vs vegetable'), and then study the effects of the variance in the environments on the animals using multiple regression techniques?

In the first chapter in this Part, Jim Flynn summarises his remarkable book *Asian Americans: Achievement beyond IQ.* Our main reason for including this is that his work highlights the importance of numerous personal and environmental variables typically neglected by psychologists.

The second chapter develops this discussion, showing first that failure to develop an alternative framework for thinking about and assessing individual differences results in widespread failure to develop and utilise human talents and, indeed, to endless unethical procedures and decisions in education and human resource management, not to mention unscientific and unethical conclusions in research.

However, since this problem was highlighted by none other than Spearman almost a century ago, one is obliged to address the question of why the topic has been neglected.

At that point, one is not merely led to consider the external social forces which, as Flynn says, so much determine behaviour, but actually to re-think the very framework psychologists deploy to try to understand behaviour. The transformation is as great as that which Newton introduced into physics. Before Newton, if things moved or changed direction it was because of their *internal* properties: they were *animated.* After Newton, it was mainly because they were acted upon by a network of *external* forces which could nevertheless be mapped, measured and harnessed.

It is argued that psychologists have not merely largely neglected these external forces, it is these forces which have mainly contributed to their failure to develop a more appropriate framework for thinking about and assessing individual differences … and this neglect is contributing in vitally important ways to the network of forces that are heading our species toward extinction at an exponentially increasing rate.

The final chapter in this Part of our book returns to the question of the problems involved in establishing the validity of a test which claims to measure meaning making ability, looks in more detail than we did in the General Introductory Chapter at the occupational predictive validity of the RPM, and briefly outlines an alternative way of thinking about individual differences grounded in the mid-career work of David McClelland.

Chapter 18

Asian Americans: Achievement Well Beyond IQ*

Jim Flynn

Abstract

The problem addressed in this paper is first to thoroughly document, and then to explain, the impressive scholastic, occupational, and income achievements of Asians in America. In the past, some psychologists have cited apparently impressive evidence of a superiority in general cognitive ability. It turns out that that this evidence is seriously flawed – and not merely because of failure to allow for the intergenerational increase in scores but also as a result of seemingly endlessly compounded sampling deficits and corrections and adjustments introduced into the norming studies. Be that as it may, with IQ held constant, the Asian's achievements exceed those of Whites by a huge amount. Once an IQ-based explanation has been discredited, attention focuses on issues rarely discussed by psychologists – such as other psychological characteristics and multiple cultural supports. These are contrasted with those operating in other cultural groups, some of which perform far below what might otherwise be expected.

Note: I hope I have excerpted enough from my book, *Asian Americans*: *Achievement beyond IQ* (Flynn, 1991), to whet appetites. But only the original provides the detail needed to support the argument.

****

Some 40 years ago, Nathanial Weyl (1966, 1969) gave Chinese and Japanese Americans a prominent place in his American natural aristocracy. He noted that Chinese Americans had three to five times their proportionate share of college faculty, architects, scientists, school

* An earlier version of this chapter has for some time been available on the Web Psych Empiricist: http://wpe.info/papers_table.html

teachers, engineers, and physicians and that Japanese Americans excelled in the same fields, although to a lesser degree, and had twice their proportionate share of artists and writers. In 1985, the upper 70% of Asian 18-year-olds took the Scholastic Aptitude Test (SAT) and matched the upper 27% of Whites (ETS, 1985, 1988). Between 1981 and 1987, Asian American high school students were much overrepresented among winners of National Merit Scholarships, U.S. Presidential Scholarships, Arts Recognition and Talent Search scholars, and Westinghouse Science Talent Search scholars. The last is America's most prestigious high school science competition and in 1986, the top five winners were all Asian Americans.

During the 1980s, there was an explosion of articles about Asian Americans in publications like *The New York Times* and *Time Magazine.* Their numbers at prestige universities had made a powerful impression on the popular imagination. Asian Americans were just over 2% of the population and yet by 1987, they were 14% of the entering class at Harvard, 16% at Stanford, 20% at MIT, 21% at Cal Tech, 25% at Berkeley. When journalists approached Arthur Jensen for an explanation, he endorsed the view that Asian Americans do so well because they are smarter, citing several IQ studies of Chinese Americans (Brand, 1987). However, the real foundation of belief in the high IQs of Chinese and Japanese Americans lay elsewhere: Vernon's great book *The Abilities and Achievements of Orientals in North America* (1982).

Vernon concluded that Chinese American's nonverbal IQ had risen from parity with Whites in 1965 to about 110, 10 points above the White average. I became suspicious when I realized that he had relied heavily on Jensen's testing of children from San Francisco's Chinatown on the Lorge-Thorndike Intelligence Test, and a study by Jensen and Reynolds (1982) of the Berkeley, California, public schools which yielded very high Lorge-Thorndike IQs. I knew from personal correspondence with Jensen that the Berkeley study had actually been done in 1968 and wondered if Vernon had thought it done circa 1980. A 10-point rise in Chinese nonverbal IQ (from 100 to 110) between 1965 and 1980 was unlikely, but such a rise between 1965 and 1968 was quite incredible. Moreover, when the elite Chinese of Berkeley were compared to the elite Whites of Berkeley, the Chinese actually had somewhat lower IQs. And the IQ values for both races looked odd. For example, Berkeley Whites had 118 for verbal IQ and 120 for nonverbal IQ: no school district in America should have an average IQ that high, however elite it might be.

I began to suspect that Vernon was misled by something unknown in his day: that massive IQ gains over time render test norms obsolete and obsolete norms inflate IQs. And I realized that if Vernon was mistaken, we needed a whole new pair of spectacles. Up to now, high IQ and high achievement seemed to reinforce one another as evidence of the superior intelligence of Chinese and Japanese Americans. But if their mean IQ were no higher than Whites, or even below Whites, then their ordinary IQs and extraordinary achievements would dictate that non-IQ factors have a potent role in group achievement. That, of course, has important implications not only for Chinese and Japanese Americans, but also for other ethnic groups and gender groups. A problem that seemed rather humdrum (they do so well because they are smarter) suddenly posed a challenge to the intellect.

Reassessing Chinese and Japanese IQ

I cannot here indicate the range of studies that had been affected by obsolescent norms but will offer one illustration. Werner, Simonian, & Smith (1968) studied all Japanese children born in 1955 on Kauai Island (the north western-most island of the Hawaiian chain). In 1965 to 1966, they gave these children, now aged 9 to 10 and numbering 253, the SRA Primary Mental Abilities Test, Elementary Form, 1954 edition, and put their mean IQ at 108. Investigation revealed that the norms against which these children were scored suffered from a total of 33.5 years of obsolescence.

Werner had not used the 1962 edition of the PMA test, but the 1954 edition, presumably because a backlog was available. The 1954 test manual and technical supplement tell a sad story. The manual (Thurstone & Thurstone, 1954a, p.1) says that the test has been "improved" by having its norms equated with those of the Stanford-Binet, which refers of course to the 1937 Stanford-Binet whose standardization sample was tested in 1932 (Flynn, 1984, p. 30). The technical supplement (Thurstone & Thurstone, 1954b, pp. 2-4) tells why. In 1951, the Thurstones found that their test was giving lower IQs than the Stanford-Binet and adjusted their norms accordingly.

Actually, assuming the PMA test was normed in 1946 (shortly before the 1948 edition), the score difference was not a product of bad sampling but of IQ gains over time. Flynn (1984, p. 35) showed that IQ gains

in America since 1932 have proceeded at a general rate of .3 points per year. The 14 years between 1932 and 1946 would mean a gain of 4.2 IQ points, and would toughen the PMA norms by that amount, which predicts almost perfectly the 4.3-point deficit that so disturbed the Thurstones (Thurstone & Thurstone, 1954b, p. 3). In 1954, rather than realizing their norms needed to be updated, they projected them even further back into the past, presenting norms that were 22 years obsolete on the day of publication. In 1965-1966, when Werner et al. scored their Japanese subjects against them, the norms were 33.5 years obsolete and inflated the IQ scores by 10.05 points (.3 points per year x 33.5 years = 10.05).

Table 18.1 gives the summary results of my reanalysis of studies of Chinese and Japanese grade and high school children. Between 1960 and 1975, these children had a mean IQ slightly below that of their White counterparts.

IQ and Occupation

The children in Table 18.1 were born predominately between 1945 and 1949. Therefore, I will focus on the achievements of Chinese and Japanese Americans born between those years. They outperformed Whites at school. Only half as many lagged a grade or more behind their age group, 95% eventually graduated from high school as compared to less than 89% of Whites. At least 50% of them took the Scholastic Aptitude Test, as compared to less than 30% of Whites, and despite this they matched White performance. They maintained the same 5 to 3 ratio when undertaking graduate study. In their early 30s, the Chinese American cohorts out-numbered Whites in high status occupations by 55% to 34%, the Japanese cohorts outnumbered Whites by 46% to 34%.

Table 18.2 uses the ratios of Chinese and Japanese Americans to whites in high status occupations to measure what I call the IQ/achievement gap. The IQ thresholds white American need to exceed to qualify for certain occupations is well documented. You find few whites in high status occupations unless their IQs are at least average (100 or above). I decided to use the superior ratios Chinese and Japanese enjoyed in those occupations to estimate what IQ a white subgroup would have to have in order to achieve such an occupational superiority. For example,

Table 18.1. Chinese and Japanese Americans: Mean IQs 1938 to 1985

Study[a]	Year	Age *n*	Age (years)	Status[b]	IQ Nonverbal	IQ Verbal	IQ Overall
Coleman Report	1965	3995	8-17	1	99	95	97
Berkeley	1968	234	9–11	2	98	95	96
Kauai	1965-66	253	9–10	2	99[c]	97[c]	98[c]
Wyoming	1943–45	669	17	2	–	96	–
Chinatown	1972	53	9	2	101[c]	91[c]	96[c]
Project Talent	1960	150	17	2	97	96	96
Chinatown	1975	254	9–11	3	101[c]	89[c]	95[c]
Hawaii	1960–63	554	15–17	3	–	96[c]	–
Los Angeles	1969–70	390	16	3	99[c]	95[c]	97[c]
Ethnic Project	1950–69	929	–	3	–	–	98
Western City	1977	98	25	4	99	92	96
Berkeley (U)	1966	309	18	4	–	94	–
S. California (U)	1975–76	42	18–19	4	102	–	–
Hawaii (U)	1978–79	118	20	4	101	98	99
Honolulu (E)	1938	3008	10–14	(1)	99	86	93
Westown (A)	1985	317	12–17	(1)	94	87	91

Note. The studies are fully described in Flynn (1991).

[a]Key to symbols: (U) = university students; (E) = subjects from an earlier period, that is, pre-war; (A) = Asian subjects from a later period with only a minority of Chinese and Japanese.

[b]Key to status categories: 1–3 denote various degrees of reliability; 4 = samples selected by criteria that render them necessarily unrepresentative; (1) = well-selected samples of groups unrepresentative of Chinese and Japanese 1945–1975.

[c]These values were substantially altered to adjust for either obsolete norms or sample bias.

Table 18.2. Chinese and Japanese Americans 1960 and 1980
Difference Between Actual IQ and IQ Estimated on the Basis of Occupation

Occupation	IQ				
	Threshold[a]	Ratio[b]	Estimated[c]	Actual[d]	Difference
		Chinese 1960			
Elite professions	110	2.264	113	99	14
Professions & technical	100	1.766	120	99	21
		Chinese 1980 (resident pre-1970)			
Elite professions	110	2.433	114	99	15
Professions & technical	100	1.793	121	99	22
Professions, technical and managerial	100	1.584	114	99	15
		Chinese 1980 (native born)			
Professions, technical and managerial	100	1.526	113	99	14
		Japanese 1960			
Elite professions	110	1.228	102	99	3
Professions & technical	100	1.299	106	99	7
		Japanese 1980 (resident pre-1970)			
Elite professions	110	1.367	103	99	4
Professions & technical	100	1.236	105	99	6
Professions, technical and managerial	100	1.161	104	99	5
		Japanese 1980 (ages 25 to 44 and resident pre-1970)			
Professions, technical and managerial	98	1.305	108	99	9
		Japanese 1980 (ages 25 to 44 and native born)			
Professions, technical and managerial	98	1.361	109	99	10

[a]The threshold gives an IQ above which about 90% of those in each occupational category would score.
[b]The ratio is the per capita ratio of Chinese or Japanese to White in each occupational category.
[c]Estimated IQ refers to means for Chinese or Japanese Americans in general calculated from the thresholds and ratios.
[d]Actual IQ is the mean for the 12th graders of the Coleman Report (1966).

if a white subgroup had a ratio of 1.793 to 1 compared to whites in general for high status occupations, it would be normal to expect them to have a mean IQ well above white average (121 rather than merely 100). The surplus 21 points stand as the IQ/achievement gap. It means that Chinese Americans could spot whites 21 IQ points and till match them for occupational status.

Table 18.2 gives IQ/achievement gaps (estimated IQ minus actual IQ) for Chinese and Japanese Americans who were aged 16 years and over at the time of the 1960 census, or at the time of the 1980 census. If we take those who had achieved the occupation of their maturity, those aged 30 years and over, the 1960 data give the occupational achievements of people who had left school before 1948. Most of those from the 1980 census, those aged 38 years and over, had left school before 1960. Yet, our estimates for the actual IQs of Chinese and Japanese come from those who were school children during the 1960s. The match between IQ and achievement is poor. In order to get a good data match between school-tested IQ and eventual adult occupations, we will follow the Coleman Report 12th graders through to their 1980 occupations, and compare *their* estimated IQs with *their* actual IQs.

Table 18.3 does this. The Coleman Report 12th graders were aged 17-18 years in 1965 and by 1980, they were aged 32-33. Therefore, Table 18.3 takes the occupational profile of ages 30-34 from the 1980 census and removes all those who arrived in America after 1965. Therefore, it at least simulates following the Coleman Report cohorts from IQ testing as high school seniors to their occupational achievements as adults. It also uses IQ thresholds appropriate to these cohorts, thresholds that take into account the increased number of young adults in high status occupations by 1980. The Chinese total cohort has an IQ/achievement gap of 21 points, one point lower for the native born, and the Japanese total cohort a gap of 10 points, one point higher for the native born. These constitute our best estimates of the Chinese and Japanese IQ/occupational achievement gaps, estimates that I sometimes round off to 20 and 10 points respectively.

IQ and Income

There is a positive correlation between IQ and income, albeit much lower than that between IQ and occupation. The purpose of the next section is to compare the actual IQs of Chinese and Japanese Americans with the estimated IQs we would posit based on their incomes.

Table 18.4 shows that the mean IQ of Chinese and Japanese Americans badly underestimates their incomes: Chinese earn almost $2,200 more than predicted and Japanese almost $1,900. Both groups have had their median incomes boosted by adjustments. The Chinese

Table 18.3. Coleman Report Cohorts: Difference Between Actual IQ As 12th Graders (1965) and IQ Estimated on the Basis of Occupation 15 Years Later (1980)

	IQ				
Group	Threshold[a]	Ratio[b]	Estimated[c]	Actual	Difference
Chinese total cohort[d]	97	1.588	120	99	21
Chinese native born[e]	97	1.572	119	99	20
Japanese total cohort[d]	97	1.323	109	99	10
Japanese native born[e]	97	1.345	110	99	11

[a]The threshold applies to Americans, aged 30 to 34 years in 1980, who were in managerial, professional, and technical occupations; it gives an IQ above which about 90% of them would score.

[b]The ratio applies to those members of the groups listed who, in 1980, were aged 30 to 34 years and in the occupational categories named; it gives the per capita ratio of Chinese or Japanese to White.

[c]Estimated IQ refers to mean IQs for the groups listed calculated from the thresholds and ratios (see below).

[d]Total cohort refers to all Chinese and Japanese Americans who were 12th graders in American high schools in 1965.

[e]Native born refers to the American born members of the total cohorts.

Example of calculations, estimated IQ of Chinese total cohort:
(1) White mean and *SD* = 100 and 15;
(2) 97.25 (IQ threshold) is .183 *SDs* (2.75 ÷ 15 = .183) below White mean;
(3) Percentage of Whites above 97.25 = 57.26%;
(4) 57.26% x 1.5876 (Chinese to White ratio) gives 90.91 as percentage of Chinese above 97.25;
(5) Chinese mean is 1.335 *SDs* above 97.25;
(6) 1.335 x 16.74 (Chinese *SD*) = 22.35 as IQ points to be added to 97.25; (7) 97.25 + 22.35 = 119.60 or 120 as estimated IQ.

went from parity with American Whites to $2,000 above, primarily after removal of post-1970 immigrants, but with a small gain from being equated with the White sex ratio. The Japanese went from $1,100 above Whites to $1,700 above, primarily because of the sex ratio factor, with some of that gain lost by adjustment for age. The removal of those who arrived between 1970 and 1980 was dictated by the IQ data available but most recent immigrants are at a great disadvantage in terms of income, so their removal has the added benefit of a fairer comparison between Chinese and Whites. The Japanese were not much affected by this because their recent immigrants have been few and elite.

Table 18.4 puts the Chinese IQ/achievement gap based on income at 16 points. This is one point higher than the estimate based on their representation in professional, technical, and managerial occupations in Table 18.2. These two estimates are the appropriate ones to compare because they cover essentially the same age groups: all those 15 years and over as compared to all those 16 and over. The Japanese estimate based on income is 13.8 IQ points. This is greater than all the occupational estimates, even the 10-point estimate from the Coleman Report cohort (Table 18.3). Japanese occupational estimates approach their income estimate only when those most affected by relocation centers are removed from the former but not the latter. Perhaps the World War II evacuation did more to restrict entry into high status occupations than it did to reduce the capacity to make money.

The Roots of "Overachevement"

Our best estimate of the size of the IQ/achievement gap is 21 IQ points for Chinese, 10 points for Japanese. I should add a qualification from the perspective of 2006. Recent studies tend to show that today's Chinese and Japanese Americans have a modest IQ advantage on whites. They, of course, are children and grandchildren of the Chinese and Japanese Americans I have analyzed, namely, those born between 1945 and 1949. The earlier generation came from homes of average socio-economic status and had average IQs or slightly below. Their high incomes and occupational status have given their offspring advantages they did not enjoy, so it is no surprise that the child has surpassed the parent for IQ. That in itself does not mean that the occupation/achievement gap is any less for today's Chinese and Japanese Americans. I leave that study to a younger scholar.

Table 18.4. Chinese, Japanese, and White Americans 1979: IQ and Income

Difference between actual income and income estimated by IQ

		Income in US dollars		
Group	IQ	Estimated	Actual[a]	Difference
White	100	15,704	–	–
Chinese	98.5	15,502	17,668	2,166
Japanese	98.5	15,502	1, 364	1,862

Difference between actual IQ and IQ estimated by income

		IQ		
Group	Income	Estimated	Actual	Difference
Chinese	103.1	114.6	98.5	16.1
Japanese	102.6	112.3	98.5	13.8
White	100.0	100.0	100.0	–

[a]Median income of full-time workers, Chinese and Japanese resident pre-1970; adjusted in terms of White sex ratio and age distribution -see text.

[b]The actual incomes have been translated into values analogous to IQ scores, that is, the mean for White income was set at 100, the SD at 15.

Examples of calculations:
Chinese estimated income:
(1) 100.0 (White mean) – 98.5 (Chinese mean) = 1.5;
(2) 1.5 ÷ 15 (White *SD*) = .10 *SDU;*
(3) .10 x .213 (path coefficient between IQ and income) = .0213 *SDU;*
(4) .0213 x $9497 (White *SD*) = $202;
(5) $15 704 (White median) – 202 = $15 502.

Chinese estimated IQ:
(1) $17 668 (Chinese median) – 15 704 (White median) = $1964;
(2) $1964 ÷ 9497 (White *SD*) = .207 *SDU* ;
(3) .207 x 15 (White *SD*) = 3.10;
(4) 3.10 ÷ .213 (path coefficient between IQ and income) = 14.6;
(5) 14.6 + 100 (White mean) = 114.6.

Whatever the exact size of these gaps, their existence shows that ignoring ethnic differences, particularly between Chinese and white Americans, does not work. A group of Whites with the same mean IQ as Chinese Americans would fall far below their achievements; a group of Whites with the same achievements as Chinese Americans would have a much higher mean IQ. The path by which Chinese overachieve compared to White Americans is clear. It begins with achievement tests at school, passes through the Scholastic Aptitude Test (SAT) and university entrance, passes through the Graduate Record Exam (GRE) and entry into graduate or professional schools, and culminates in their high occupational profile. But what factors lie behind that path and play of role of cause?

I am not going to explore the possibility that Chinese have a genetic superiority for IQ. Certainly, their IQs in American do not signal such. It might be argued that those who came to America prior to 1950 were substandard compared to those who remained at home. The facts call this into question. The earlier immigrants were unskilled labourers and few of the students who came to study in America were allowed to remain. The Chinese who became American citizens brought over their own children and, after 1924, these were the main source of immigration. Many of these children were fictitious products of the "slot racket" and insured a fairly random sample of the home village. For example, a Chinese born in America in 1870 had by 1957 brought over 57 of his "sons", who had in turn brought over 250 of their "sons", which is to say he was personally responsible for the entry of almost the entire male population of his village. In 1957, it was estimated that at least half of San Francisco's Chinatown were products of the slot racket or other forms of illegal immigration (Lee, 1960, pp. 78-81, 95, & 302-304; Wang, 1966, pp. 96-98).

If cultural differences are the root cause of Chinese overachievement, let us consider what those might be. The analysis here has more in common with the multi-component historical analyses of eg Braudel (eg 1993) than with those who have searched for "basic" variables, such as, among sociologists, Weber (1930) or, among psychologists, McClelland (1961). I will focus on the origins of Chinese, Irish, and Black Americans.

Three Histories

The Chinese who came to America before 1950 came predominately from the Pearl River delta. This area has been the home of an intense rice-based agriculture for over 4,000 years. The unrelenting work demanded may be greater than any other area in the world. Two rice crops and one

dry crop are produced each year. Horticulture produces vast quantities of fruit, tea, and silk (from mulberry bushes) are marketed, vegetables and sweet potatoes grown, livestock include chickens, pigs, buffaloes, and fish farms. These conditions engendered a powerful work ethic and Chinese immigrants to America have manifested that ethic from the 1850s right up to the present. Every observer has commented on the pace of work, the hours of work, the propensity to save and invest in their children's education. Lee notes something that adds a fascinating corollary to our thesis. The Sze Yap people have been less achieving than other Chinese Americans: these people came from the periphery of the Delta where soil was less fertile and agriculture less intense (Bodde, 1957, p. 52; Brand, 1987; Butterfield, 1990; Fairbank, Reischaur, & Craig, 1965, pp. 90-91; Lee, 1960, pp. 52, 144-145, & 254-257; Petersen, 1978, p. 75; Tan, 1986, pp. 16-17 & 171-171; Vernon, 1982, pp. 274-275).

Irish immigrants came from a 19th-century rural Ireland in which conditions could not have been more different. Half the rural population lived in mud huts, tilling a quarter to a half-acre farms only one-sixth the size of those prevalent in China. In order to avoid starvation, these farms were given over almost entirely to the optimum crop, namely, potatoes. Potatoes required little more than spading and turning a few weeks of the year. All improvements were the property of the landlord, and tenants could be turned out at will. Irish peasants spent most of the year in enforced idleness. They were not crushed. Travelers remarked on their hospitality, love of music and dance, and gaiety. But no potent work ethic developed. When the Irish came to America, they were often content with a bare sustenance, even this was a welcome relief after famine Ireland. They made a grand thing out of Saturday night, given over to sociability and fighting, and if the best street-fighter on the block died poor, he had moments of glory unknown to a cost accountant. Irish Americans may have lacked a positive attitude toward work but some of them had a very good time. (Glazer & Moynihan, 1970, pp. 238-239, 246, & 259-262; Lee, 1960, pp. 385-386; McAleavy, 1967, p. 31; Woodham-Smith, 1962, pp. 18-37, 268, & 409).

Traditional China gave education an all-pervasive role, indeed, it provided the foundation on which rested the entire political, social, economic, and cultural life of the Chinese people. Confucianism conferred dignity on peasant labor, peasants were ranked second only to the Mandarin class, and the traditional Chinese examination system was the only way a village youth could rise to the Mandarin class. The

periodic examinations were great public events and preparation for them so arduous that it led to a virtual examination way of life. Those who passed the first level were called "budding geniuses", those who passed the second "promoted scholars", those who passed the third became high officials, the best often became prime minister and married a royal princess. Those who attained high office were expected to foster the interests of their villages of origin and whole families, clans, and villages pooled their resources to give their brightest boy the leisure to prepare for the exams (Fairbank, Reischaur, & Craig, 1965, pp. 84-88; Hu, 1962, pp. 3 & 13-15; Lee, 1960, pp. 96-97; Menzel, 1963; Wang, 1966, pp. 13-14).

American Chinese from the start emphasized education and looked upon money earned from academic status and professional credentials as more honorable than mere money alone. The Chinese family became one of the most educationally efficient in America, rivaled only by the Japanese and Jews. There was the usual generational strife found in immigrant families, but above the battle certain assumptions were rarely contested. Children were expected to study hard and did so, earning high marks irrespective of IQ, which gave the Chinese unusually low IQ thresholds for entry into high status occupations. The Coleman Report shows Oriental students doing many more hours of homework, having better attendance records and higher aspirations; the National Longitudinal Study adds confirmation, plus showing they spent far less time on athletics and extra-curricular activities. Parents tried to protect their children's time by discouraging part-time jobs.

Chinese youths identified their self-esteem with academic advancement, targeted themselves for the best universities, and rarely passed up a chance for professional status when they could qualify, which gave Chinese Americans as a group a high capitalization rate on their available pool of talent. An Irish youth might forfeit a promising opportunity so as to attend a Catholic college, stay with kin or friends, marry the girl or boy next door, a Chinese rarely (Coleman et al., 1966, p. 24; Hsia, 1988, p. 78; Lee, 1960, pp. 185-230, 374, 382, & 392; Petersen, 1978, pp. 92-93; Rock et al., 1985).

Ireland was the only European country that did not establish a single university during the Middle Ages. By the 19th century, the mass of people had no educational tradition of any sort thanks to 130 years of the penal laws which forbade Catholics from attending school, running schools, even sending their children abroad to be educated. Even those

few who escaped to Britain and whose children attended college were remarked upon by their contemporaries for their lack of commitment.

When Irish immigrants came to America, there was no presumption that families should sacrifice to educate their young. The first objective was ownership of a family home and everyone was expected to contribute: children dropped out of school, sacrificing education and future skills, to work and augment family capital and income. Devout parents discouraged education as a threat to faith. When the Irish rose out of poverty, they did not identify their worth with professional advancement, but sought status as political orators, singers, entertainers, athletes, military heroes. For many Irish, the ideal was a secure civil service job and real life was lived outside of work, arguing religion or politics or becoming the best raconteur at the local saloon (Kessler-Harris & Yans-McLaughlin, 1978, pp. 114-120; Glazer & Moynihan, 1970, p. 258; Sowell, 1975, pp. 71-80, 127, 146-147, & 205; Woodham-Smith, 1962, p. 27).

Cotton plantation Blacks came from a slavery and peonage at least as devoid of self-motivated work and educational tradition as the Irish, and arrived at urban centers two or three generations later. Certain Irish institutions counterproductive vis-à-vis the Chinese take on the character of priceless assets when viewed against the backdrop of the Black experience. The Catholic Church with its parochial schools and universities may have given the Irish an education mediocre by comparison with the Chinese, for the latter extracted the best education public schools and great universities had to offer. But Catholic schools gave Irish the literacy and numeracy that led toward the middle class, whereas Blacks faced the worst schools the public system offered. The Irish political machines may have encouraged them to be content with modest civil service jobs and discouraged higher ambitions. If so, the numerical and political weakness of the Chinese removed a temptation and encouraged them to rely on that combination of hard work, sobriety, and maximization of educational capital, which eventually led them to the pinnacle of achievement in American society. On the other hand, political patronage did lift many Irish into the middle class, favoring them over Blacks who became politically dominant only later, their freedom of maneuver limited by entrenched groups and a climate unfavorable to patronage.

Despite these relative disadvantages, as late as 1970 it seemed Blacks could hope to follow the Irish path toward parity. However, the years between 1965 and 1990 saw the development of a trend that threatened to divide Black America into a middle class showing excellent

progress in terms of enhanced occupational status, and a lower class whose family structure was becoming less and less educationally efficient. The trend was toward fewer Black males in steady employment. The causes included new conditions in the labor market, thanks to regulation and a shift in the locus of unskilled jobs, and the rise of an alternative economy based on drugs. Other causes as yet unknown are probably operative: the relative success of the children of free persons of color and West Indian immigrants suggests that these are environmental rather than genetic.

A Last Word

The powerful emotions engendered by group differences in test scores, academic achievement, occupation, and income take place in a certain context. That context is the product of human misery. If America can help almost all of its citizens toward a good life, the obsession with total equality will diminish: whether Chinese or Irish or Blacks have exactly the same occupational profile may still interest social scientists but not the ordinary person.

References

Bodde, D. (1957). *China's cultural tradition*: *What and whither?* New York: Rinehart.

Brand, D. (1987). The new whiz kids: Why Asian Americans are doing so well, and what it costs them. *Time*, August 31.

Braudel, F. (1993). *A History of Civilizations.* London: Penguin Books.

Butterfield, F. (1990). Why they excel. *Parade*, January 21.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Hood, A. M., Weinfeld, L. D., & York, R. L. (1966). *Equality of Educational Opportunity*. Washington, DC: U.S. Government Printing Office.

Educational Testing Service (1985). *College bound seniors.* Princeton, NJ: College Entrance Examination Board.

Fairbank, J. K., Reischaur, E. V., & Craig, A. M. (1965). *East Asia*: *The Modern Transformation.* Boston: Houghton Mifflin.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29-51.

Flynn, J. R. (1991). *Asian Americans*: *Achievement Beyond IQ*. Hillside, NJ: Lawrence Erlbaum Associates.

Glazer, N., & Moynihan, D.P. (1970). *Beyond the melting pot* (2nd ed.). Cambridge, MA: MIT Press.

Hsia, J. (1988). *Asian Americans in higher education and at work.* Hillsdale, NJ: Lawrence Erlbaum & Associates.

Hu, C-T. (1962). *Chinese education under Communism.* New York: Teachers College, Columbia University.

Jensen, A. R., & Reynolds, C. R. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences, 3*, 423-438.

Kessler-Harris, A., & Yans-McLaughlin, V. (1978). European immigrant groups. In T. Sowell (Ed.), *Essays and data on American ethnic groups* (pp. 107-137). Washington, DC: The Urban Institute.

Lee, R. H. (1960). *The Chinese in the United States of America.* Hong Kong: University Press.

McAleavy, H. (1967). *The modern history of China.* London: Weidenfeld & Nicolson.

McClelland, D. C. (1961). *The Achieving Society*. New York: Van Nostrand.

Menzel, J. M. (1963). Introduction. In J. M. Menzel (Ed.), *The Chinese civil service*: *Career open to talent?* (pp. vii-xii). Lexington, MA: Heath.

Petersen, W. (1978). Chinese Americans and Japanese Americans. In T. Sowell (Ed.), *Essays and data on American ethnic groups* (pp. 65-106). Washington, DC: The Urban Institute.

Rock, D. A., Ekstrom, R. B., Goetz, M. E., Hilton, T. L., & Pollack, J. (1985). *Contractor report*: *Factors associated with decline of test scores of high school seniors, 1972 to 1980*; *a study of excellence in high school education, educational policies, school quality, and school outcomes.* Princeton, NJ: Educational Testing Service.

Sowell, T. (1975). *Race and economics.* New York: Longman.

Tan, T. T. (1986). *Your Chinese roots*: *The overseas Chinese story.* Singapore: Times Book International.

Thurstone, L. L., & Thurstone, G. T. (1954a). *SRA Primary Mental Abilities, examiner manual, ages 7 to 11* (2nd ed.). Chicago: Science Research Associates.

Thurstone, L. L., & Thurstone, G. T. (1954b). *SRA Primary Mental Abilities, technical supplement, ages 7 to 11* (1st ed.). Chicago: Science Research Associates.

Vernon, P. E. (1982). *The abilities and achievements of Orientals in North America.* New York: Academic Press.

Wang, Y. C. (1966). *Chinese intellectuals and the West, 1872-1949.* Chapel Hill, NC: University of North Carolina Press.

Weber, M. (1930). *The Protestant Ethic and the Spirit of Capitalism*. New York: Scribner.

Werner, E. E., Simonian, K., & Smith, R. S. (1968). Ethnic and socioeconomic status differences in abilities and achievement among preschool and school-age children in Hawaii. *The Journal of Social Psychology, 75*, 43-59.

Weyl, N. (1966). *The creative elite in America.* Washington, DC: Public Affairs Press.

Weyl, N. (1969). Some comparative performance indexes of American ethnic minorities. *Mankind Quarterly, 9*, 106-119.

Woodham-Smith, C. (1962). *The great hunger.* London: Hamish Hamilton.

Chapter 19

Intelligence, Engineered Invisibility, and the Destruction of Life on Earth*

John Raven

Abstract

Gottfredson (1997) assembled a huge amount of data supporting three main claims: Out of all the traits known to psychology, only ***g*** predicts much of the variance in occupational performance; ***g*** is the most important of all the variables assessable by psychologists determining the effectiveness of behaviour outside work; and occupational status depends mainly on ***g***. In this article it is shown that, in both the workplace and educational system, other qualities besides ***g*** *are* important but remain invisible. This invisibility is produced by a network of interacting, but mutually supportive, processes which include the adoption of an inappropriate psychometric model and limited criteria of performance, but, most importantly, from what seems to be a sociological "need" for a single and unarguable criterion of merit to legitimise a social hierarchy which contributes enormously to the network of forces which result in most people spending most of their time contributing to activities which are, directly or indirectly, destructive of other people's quality of life and the chances of our species and the planet surviving – that is, to activities which can only be regarded as highly unethical. Embracing the task of mapping these socio-cybernetic forces results in focussing on the external rather than the internal determinants of behaviour. Trying to map these forces has enabled us to outline arrangements which should make it possible to run the educational system–and other domains of human

* An earlier version of this paper has for some time been available at WebPsychEmpiricist: http://www.wpe.onfo/papers_table.html

endeavour–more effectively. These developments depend quintessentially on organisational arrangements, job descriptions, and appraisal systems the development of which falls clearly within the domain of organisational psychology.

Overview

In a wonderfully documented paper, Gottfredson (1997) not only argues that ***g*** is the major variable responsible for differential performance in all walks of life (or at least the only one whose contribution can be demonstrated with the assessment instruments currently available to us) but also the main factor lying behind our hierarchical social order.

In this paper, it is first shown that, at least in the workplace and the educational system, numerous other qualities *are* important but remain invisible if one utilises only tools developed within the traditional measurement paradigm, focuses mainly on conventional criteria of job performance, and accepts assumptions about the functionality of hierarchical organisation of workplaces and society.

Next, it is argued that all of these things–failure to recognise, develop, and utilise the wide range of hidden talents that are actually available, the dominant criteria of job-performance, and our hierarchical social order–are seriously socially dysfunctional in the short term and, more especially, in the longer term. Nurturing the competence to understand and intervene in the networks of invisible social forces that overwhelmingly determine our individual and collective behaviour is therefore an activity of the greatest importance.

More than that, from a scientific point of view, it is vital to develop more systematic ways of illuminating and intervening in such networks of forces.

Our studies of the educational system are used to illustrate how this can be done. It is shown that such networks of forces, better termed "socio-cybernetic systems", can be exposed by using psychological data to illuminate the hidden social processes that are at work.

What then emerges is that these neglected *external* forces are among the most important determinants of behaviour. To move forward in psychology, we need a paradigm shift as great as the Newtonian shift from attributing movement to the internal, "animistic", properties of moving bodies to accounting for it largely by reference to networks of invisible external forces.

Finally, it is shown that even a preliminary understanding of the socio-cybernetic forces controlling the operation of the educational system enables us to design an alternative management system which would make it possible to run the system in such a way as to achieve its manifest goals more effectively. The requisite design would in fact be almost the exact opposite of that which informs most of the policies currently being pursued worldwide. Implementing that design is crucially dependent on psychologists developing new specifications for the requisite organisational arrangements, new job descriptions, and new tools for organisational and staff-appraisal.

Part I: Other Qualities Are Important

Evidence From the Workplace

Gottfredson's first claim–i.e., that ***g*** and not much else has predictive validity in occupational setting–is well supported by data brought together by such authors as Schmidt and Hunter (1998), Jensen (1998), and Ree, Earles, and Teachout (1994). Nevertheless, hugely impressive though these assembled data are, they are not entirely convincing.

One reason for this is that much depends on job definitions and performance appraisal systems which overlook many important contributions.

As argued in greater detail elsewhere (e.g., J. Raven, J. C. Raven, & Court, 1998), there is enormous tension between the assumed job definitions put forward in, for example, the writings of Jaques (1976, 1989) and the results of more empirical studies of the kind published by Kanter (1985), L. M. Spencer and S. M. Spencer (1993), Huff, Lake, and Schaalman (1982), Desjardins and Huff (2001), Schön (1973, 1983, 1987), Russ-Eft and Brennan (2001), and the author (Raven, 1997).

In the course of hundreds of studies using fine-grained methodology – and especially *Behavioural Event Interviewing* (a variant of Flanagan's *Critical Incident Technique*)–it has been shown that effective organisations call on even their "low-level" employees (lavatory attendants, machine operatives, sales people, etc.) to utilise high-level competencies. For example, a compilation of "effective" behaviours observed among machine operatives included examples of them studying the way the overall system of which they formed a part functioned and working out for themselves what they should be doing–and doing it without having to

be given instructions. But, as researchers like Kanter (1985) and Schon (1983) have shown, even this behaviour is gross compared with the diverse subtle contributions that people in effective organisations make to the emergent properties of problem-identification-and-solving networks which, while crucial to the improvement and survival of the product, services, and organisation itself, are rarely discussed.

Yet this fact cannot show up in studies grounded in correlations between psychological tests and job performance. This comes about, in part, because the classical test armoury, as a result of the psychometric model adopted, contains no good measures of the relevant qualities, and, in part, because the criteria of occupational performance adopted in virtually all these studies leave much to be desired: If, as is often the case, managers and supervisors believe that the jobs of "low level" employees involve following rules without thinking it creates a self-fulfilling propensity so that other features of performance are unlikely to show up in the studies those managers commission. What is more, as will be discussed more fully below, people's contributions are very much determined by what others do and the effects of their actions are absorbed into group processes. These contributions and effects cannot be easily identified using conventional methodology such as performance ratings.

Even *Behavioural Event Interviewing*, despite the great service it has done us by drawing attention to the importance of a huge range of occupational competencies overlooked by most researchers, often fails, because of the culturally determined associations that are evoked when one is asked to describe incidents in which people have been observed doing something considered to be particularly effective, to reveal the full extent of such contributions. As Adams and Burgess (1989) have shown in their work in schools, these associations make it unlikely that people will record incidents in which others did such things as resolve "personal" conflicts between colleagues or wrestle in private with a conceptual problem and then pass the solution on to someone who would do something about it. Yet, as both they and Kanter have shown, such activities are crucial to the creation of the cultures of intelligence or enterprise that are required for organisational survival and development.

Evidence From the Educational System

So far, I have dwelt on doubts about the validity of the "***g*** and not much else" thesis raised by our work in the occupational area. Equally serious doubts stem from our work in schools (Raven, 1977, 1985). In the course of this research (which was carried out in both elementary and secondary

schools) we observed that, as described below, when teachers set out to nurture high-level competencies through inter-disciplinary, enquiry-oriented, group-based, project work conducted in the environment around the school, huge numbers of talents, at best only marginally related to ***g***, come to light.

To give but one example: in one elementary school (Raven, Johnstone, & Varley, 1985), we found the pupils engaged in a project designed to get something done about the pollution in the local river. The project, its organisation, its effects, and the problems it posed for evaluation all merit detailed discussion, but only the briefest account can be given here. Interested readers should refer to one of the sources cited. Some pupils decided that the first thing to do was to measure the pollution in the river. Some of them then set about collecting samples of the river water and trying to analyse it. This took them to the not-so-local university where they worked with the lecturers. Note that these pupils were developing the *competencies* of the scientist: The ability to identify problems, the ability to invent ways of investigating them, the ability to obtain help, the ability to familiarise themselves with a new field, and the ability to find ways of summarising information. Other pupils decided that more progress was to be made by studying the dead fish and plants along the river bank. Still others argued that all this was beside the point: The river was clearly polluted and the problem was to get something done about it. Some then set about drawing pictures of the dead fish and plants with a view to releasing community action. The objective was not to depict what was seen *accurately,* but to represent what had been seen in a way that would evoke emotions that would lead to action. While the "scientists" mentioned above sought to describe the results of their work in what might be termed a classic academic format, other pupils again argued that that was irrelevant and set about generating slogans, prose, and poetry that would evoke emotions that would lead to outrage and action. The *criteria* for what constituted effective reading and writing thus differed markedly from those which dominate most classrooms and they varied from pupil to pupil. Still other pupils argued that, if anything was to be done about the river, it was necessary to get the environmental standards officer to do his job. (It turned out that he knew all about the pollution but had done nothing about it.) This led some pupils to set up domino-like chains to influence politicians and public servants. This in turn led the factory that was causing the problem to get at the pupils' parents saying that, unless this teacher and her class was stopped, they

would all loose their jobs. Unabashed, some pupils set about examining the economic basis for the factory's claims.

Note that this teacher was not so much concerned with enhancing pupils' specialist *knowledge* in each of these areas (though, even if it had been, documenting that knowledge would have posed insuperable problems for evaluators steeped in classical measurement theory because the knowledge to be documented was largely idiosyncratic and tacit[1]) but to nurture a wide range of *different* ***competencies*** in her pupils. These competencies were not limited to substantive areas of investigation but also included the ability to contribute to group processes, including such things as the ability to put people at ease, the ability to de-fuse the intolerance which develops between people who contribute in very different ways to a group process (e.g., the intolerance of the "artists" for the "scientists"), the ability to publicise the observations of the quiet "ideas person", and the ability to "sell" the benefits of the unusual educational process to parents. The teacher in fact devoted considerable attention to highlighting the different types of contribution which different children were making to the group process. As a result, they stopped thinking of each other in terms of "smart vs. dumb" and instead noted what each was good at.

It is extremely important to note that what was happening here involved making *descriptive statements* about each individual pupil's talents and areas of knowledge and expertise. Despite the assumptions which many of those who have grown up in the current climate of assessment bring with them, this could not be achieved by trying to arrange them on scales because a *different* set of scales would be required to record the talents of each child. To help readers get the point, it might be useful for them to try to imagine seeking to describe chemical substances in terms of profiles of ratings across each of the 96 elements. Huge amounts of useless information would be generated and the process would still fail to reveal the emergent properties that occur when different elements combine. It is what people are good at, and their idiosyncratic expert knowledge (mostly non-verbalised and consisting of knowing-how rather than knowing-that) that we need to record. To do this effectively we will need to develop a framework of agreed descriptors akin to that employed in chemistry. So far, as will be seen in the next section of this paper, all we have is a basis on which such a framework could be built.

Particularly in an American context, it is important to emphasise that the work just described, while superficially similar to the work reported

in the hundreds of accounts of project-based education that are to be found in the literature (reviewed in Raven, 1994), was in fact dramatically different to most of them because the notion of *what was to be* ***learnt*** was different. Pupils were to learn to lead, to invent, to put people at ease, to create political turbulence, etc. The objective was not that they should "learn" in the sense of acquiring stocks of standard, formal, low-level, verbal *knowledge.* The ability to build up idiosyncratic combinations of up-to-date specialist knowledge–yes–but that was different. The dozens of projects of this sort studied by Grannis (1983) and ourselves thus went far beyond those described in the widely publicised work of Gardner and his colleagues (Gardner 1987, 1991; Hatch & Gardner, 1990; Krechevsky & Gardner, 1990). The teachers we are talking about here were not dealing with six or seven "intelligences" or areas of skilled performance but with the ability to carry out one or another of a huge range of necessary, and mutually supportive, activities. It is true that all of these demand and reveal some form of intelligence and related abilities of the kind indicated by such terms as "the ability to observe" and "the ability to reason". But they also demand a wide range of additional components of competence–the ability to learn from the effects of one's actions and modify one's behaviour accordingly, the ability to persist, the ability to get help, and so on. It is also vitally important to note that none of these components of competence can be meaningfully developed or assessed generically–across all kinds of potentially valued activity–but only in the context of the specific activity being undertaken. Thus one person will display a great deal of creativity while creating classroom disruption, another while putting people at ease, and another while finding ways to undertake a scientific study. And none of them can be meaningfully assessed by asking those concerned to construct something "creative" out of a collection of toy bricks.

Conclusion to Part I

It follows from the material briefly summarised here that other qualities besides ***g*** *are* vitally important–a conclusion which in no way contradicts Gottfedson's main thesis, although it does undermine the second half of the statement "***g*** and not much else". The question the data pose is "Why, under the circumstances, has so much weight been placed on ***g*** alone in schools, workplaces, and society?"

In fact, the data so far presented go some way toward answering that question: To capture these other qualities it would be necessary to develop a very different way of thinking about and assessing human talents.

As it happens, Spearman (1926) had noted both the problem and the direction in which its solution should be sought almost a century ago. He noted that "Every normal man, woman, and child is … a genius at something … It remains to discover at what … This must be a most difficult matter, owing to the very fact that it occurs in only a minute proportion of all possible abilities. It certainly cannot be detected by any of the testing procedures at present in current usage." He also noted, first, that the ***g*** for which he is famous (and which lies at the heart of Gottfredson's thesis) had emerged from the correlations between tests that lacked both construct and predictive validity[2]. The low reliability of the tests used in the educational system is well known (see e.g. Black, 1998; Spencer, E. 1983) as is their inability to predict anything much outside the educational system (see e.g., Schmidt & Hunter, 1998). But the point being made by Spearman and the author (e.g. Raven, 1991) is more basic. The tests lack *construct* validity. There is, for example, no sense in which the typical "science" test used in schools assesses the competencies of the scientists observed in the project work discussed above or even testifies to knowledge of a sample of the "basic" information and procedures that constitute the domain of "science". The second thing that Spearman noted was that the "educational" system itself, as it actually operates, rests on a fraudulent claim because the word "education" comes from the Latin root "educere", meaning "to draw out or to develop from latent or rudimentary existence", thus implying the nurturance of diversity. If it does not mean "to put in", its outcomes cannot meaningfully be measured using tests of the kind that are most widely employed.

Given that both the multiple talent problem and the route toward its solution were noted so long ago one is forced to look for some explanation of why so little has happened. Much of the remainder of this article will be devoted to so doing.

Part II: Ways in Which Widely Accepted Assumptions in Psychology Contribute to Invisibility

1. There are basic flaws in the dominant measurement paradigm in psychology and the requisite psychometric model is at loggerheads with classical test theory.

It is easiest to illustrate some of the problems which assessment of the qualities discussed above pose for the classic assessment paradigm by reviewing the psychological nature of qualities like the ability to reason, take initiative, and work effectively with others.

All of these are difficult and demanding activities which people will neither develop nor display unless they are engaged in tasks they care about. Furthermore, "the same" activity looks very different in different contexts–just as copper looks very different when combined only with oxygen and when combined also with sulphur. Can one doubt that those whose task it is to remove dents from damaged cars "think" about what they are doing and learn from the effects of their actions even though that thinking and learning would not show up on conventional "intelligence" tests? Yet, as Spearman noted, the number of things that one person or another is strongly motivated to carry out is legion. Different people are preoccupied with "thinking", usually non-verbally, about very different things. On the basis of the limited evidence already reviewed, it also seems a reasonable hypothesis that all are creative while carrying out activities they care about–whether those activities involve creating social disruption, crafting a beautiful vase, developing a new scientific theory, establishing a harmonious personal relationship, or anything else.

If one can generalise from these examples (and evidence suggesting that one can so generalise has been brought together in Raven [1984/1997]), it would seem that constructs like the ability to reason, self-confidence, and creativity – which psychologists have for more than a century sought ways of measuring–cannot be meaningfully assessed in the way the currently dominant paradigm suggests, that is, they cannot be "measured" by presenting everyone with a common task and seeing how "well" they do, because this will fail to tap and unleash *most* people's ability to do these things. As shown in more detail in Raven and Stephenson (2001), to "measure" them one must *first* find out what it is that the individual is strongly motivated to do and then find out *which* of a number of cumulative and substitutable components of competence that they could, from a theoretical point of view, bring to bear to carry

out that activity effectively they do in fact exercise. As it happens, a procedure that operationalises this model was developed by McClelland and his colleagues in the middle of the last century (McClelland, Atkinson, Clark, & Lowell, 1958; but see Raven & Stephenson, 2001 for a re-interpretation). Those being assessed were asked to make up stories about what was happening in ambiguous pictures. They were asked to say what each of the characters in the pictures was thinking, feeling, and doing and what the outcome would be. To score the stories the psychologists concerned first asked "With what kind of activity does the person who wrote this story seem to be preoccupied?" Then, in relation to this kind of activity, and only in relation to this kind of activity, they then counted up how many of a number of different kinds of action that would potentially enable someone to undertake his or her chosen activities effectively the author imagined his characters displaying: Did they turn their emotions into the task? Did they make plans, anticipate obstacles, and seek ways of tackling those obstacles? Did they seek the help of others? Did they persist for a long period of time? It is vital to note what is going on here. This is no internal-consistency-based measure of "achievement motivation". Rather the resulting score is more like a multiple regression coefficient predicting the probability of success in undertaking an intrinsically motivating activity (with each of the individual predictors assigned the same weight). Unfortunately, even those who developed this scoring system did not recognise that what they had stumbled upon was, in reality, a radically new paradigm for the assessment of competence. Instead they presented it as a means of assessing "personality"[3]. Thereafter, in the half century that has intervened, McClelland and his disciples, in their quest for acceptability among their peers and a market for their products, largely abandoned it and came to accept and promote the classical measurement framework (see Raven & Stephenson, 2001, for a discussion of this process).

2. Problems with the accepted procedures for test validation.

When one turns to popular notions about the procedures that are appropriate for test validation, one encounters similar problems. In the workplace, people are not usually doing what other people think they are doing. As a whole series of studies, some of which have been brought together in Raven & Stephenson (2001) and Spencer and Spencer (1993) have shown, one manager is preoccupied with advancing himself in his career by running a "lean, mean" organisation and getting rid of all

the staff who would create a future, another with creating network-based working relationships which lead to the evolution of new products, another with enhancing the short-term value of the company by manipulating its image on the stock market, and so on.

Thus, to find out whether a test that claims to measure a quality like "the ability to think" does in fact do so, one cannot use criteria like supervisor's ratings or productivity (which is, in any case, a *group* rather than individual characteristic). In other words, as Messick (1995) and Raven et al. (1998) have argued, one cannot "validate" tests in the manner prescribed in most textbooks. One has, somehow, to get inside people's heads and find what they "think" *about* before one can make any meaningful statement about how well they can think.

In short, not only is the classic, internal-consistency based, measurement paradigm incompatible with the psychological nature of the qualities we have seen to be so important, so are the procedures conventionally prescribed for the validation of tests.

3. Psychologists have failed to study the emergent properties of groups.

It is widely accepted in throw-away comments made by psychologists that people are unable to function without a context and that their apparent characteristics, their behaviour, and the effects of their actions are very much determined by the context in which they live and work. Yet Kanter's (1985) research is almost alone in enquiring into these things. It emerges that the development and survival of organisations is almost entirely dependent on what takes place in what Kanter terms "*parallel organisation*" activity. During the time devoted to such activity, people come together in networks of fluid groups in which they function in ways that are *not* included in their job descriptions, deploy talents that are typically invisible and overlooked as they perform their day-to-day jobs, and adopt working practices in which the hierarchical distinctions characteristic of the day-to-day operation of the organisation are rendered inoperative. It is these invisible and idiosyncratic contributions to such emergent properties of groups as might be referred to as "climates of enterprise" that are important. How can such observations not undermine the seeming strength of Gottfredson's thesis?

The implications of these oversights can be made more obvious by drawing an analogy from chemistry. We may first ask: Where would chemists (or biologists) have got to if they had sought to describe all the variance in chemical substances (or species) in terms of one, two, five, or

even sixteen "variables"? Having come to terms with the answer to that question, we may note that the properties of copper sulphate cannot be predicted by adding the individual properties of copper, sulphur, and oxygen, and those three substances are not recognisably "the same" when studied in combination and when considered individually. Clearly, we have been headed down a blind alley. What we need is a *descriptive,* combinatorial, interaction-with-the-environment, model: A model akin to atomic theory in chemistry or to a biological classification accompanied by an account of ecological interactions and feedbacks.

4. Psychologists have accepted a great deal of sloppy thinking about "scientific methodology".

One network of beliefs supporting the hegemony of a measurement paradigm that renders many important human qualities invisible is associated with the concepts of "objectivity" that inform the recommendations of such professional bodies as the Joint Committee on *Standards for the Evaluation of Educational Policies and Programs* (1981). This particular committee recommended that only reliable and valid tests should be used in the evaluation of people and programmes. Apparently reasonable though this recommendation is, its *effect* is to render many important personal qualities and the effects of policies and educational and social activities invisible. Since there are no good measures of the main objectives and outcomes of the kind of interdisciplinary, competency-oriented, enquiry-based, education discussed earlier, the requirement that only reliable and valid tests be used in its evaluation induces researchers to use only *irrelevant* tests. This not only renders the positive outcomes of these activities invisible, it also ensures that many negative effects of conventional educational activities go undetected and undiscussed–indeed almost undiscussable[4].

The overall effect of this process is to undermine any claim to objectivity or scientific competence on the part of those concerned. In reality, such evaluations–whose main fault is a lack of *comprehensiveness*–must be considered, not only incompetent and lacking in objectivity, but also unethical. This is in part because they contribute to the process whereby most children's talents are rendered invisible and undeveloped thus contributing to the processes through which schools damage most children and their future lives and careers. But most importantly it is because these neglected talents are the very talents that are required to transform the way we live in such a way that our species–and indeed the planet as a whole–will have a chance of survival.

Part III: Ways in which Invisibility is Driven by Wider Social Processes

Having examined the contributions to the invisibility of many important components of competence that stem from assumptions or axioms arising within the discipline of psychology itself, we now consider the role of some wider social processes that seem to at work.

1. Processes operating in schools.

Most educational activity of the kind discussed has been purged from schools in the UK. This has been achieved by insisting that all pupils follow the "national curriculum" (which, in many areas, prescribes the activities teachers are to undertake on a week by week, and sometimes minute by minute, basis) and take the same tests. This has the effect of inducing teachers to pay more attention to the prescriptions of authority than to the needs of their pupils, thus rending the other talents their pupils possess even more invisible.

Our question here is: "What has driven this process?"

The most common justifications have to do with such things as eradicating "poor" teachers, facilitating the movement of pupils between schools, and improving the fairness of the procedures used to allocate position and status in a meritocracy.

But more disturbing reasons come to light as one reviews what the educational system actually does. The very least one can say is that – unlike the interdisciplinary, competency-oriented, project work discussed earlier – most of what happens in most schools amounts to a charade of little developmental or societal value (whilst conferring enormous social benefits on a minority of the participants and assigning others to lives of degradation and humiliation at the hands of the welfare "services"). This is revealed by five related observations[5]:

a) The activities which dominate most schools have little developmental value (see e.g. for schools Goodlad, 1983; Raven, 1994, for universities Murphy, 1993; Steiner, 1999);
b) Most of the tests that are used to evaluate educational performance testify to little of merit because they lack construct validity (see above);
c) These tests have little predictive validity outside the "educational" system (see Schmidt and Hunter's 1998 meta-analyses);

d) What is learned in schools rarely helps people to cope with their jobs and lives (see e.g. Bachman, O'Malley, & Johnston, 1978; Flanagan, 1978);
e) Most of the knowledge that is so painfully communicated and tested in schools is out of date when it is taught, does not relate to the problems people will have to tackle when they leave school, and, since knowledge has a half-life of a year, will be forgotten by the time it is needed (Raven, 1994).

Nevertheless, many authors have suggested that schools, while failing to nurture competence, may nevertheless, especially through the "hidden curriculum", teach some sociologically very important lessons. For example, Goodman (1962) suggested that they may teach pupils to be subservient to authority and to be willing to accept that those in higher positions know more about issues of concern than they do. Willis (1977) assembled data supporting the hypothesis that one of the primary functions of schools is to inculcate a willingness to tolerate boring work. And the author (see Raven & Stephenson, 2001) has suggested that the only conclusion one can reasonably draw from the available literature is that the system teaches people that others have the right to define who one is, what one is good at and bad at, and to allocate one's position and status on the basis of criteria they have determined. They may also teach people to abdicate responsibility for taking control over their own lives and for trying to influence what happens in society.

Still others have suggested that some of the things that are done by schools have direct sociological effects. Thus Jencks et al. (1973), in addition to producing considerable evidence supporting the view that one of the major functions of the system is to "legitimise the rationing of privilege" (by promulgating the myth that those who are advanced in it are the most "able" whilst those who are demoted are "unemployable"), also showed that its main function was to sort people into a social hierarchy. Observation of the de-humanising treatment meted out to those who fail to compete in this norm-referenced hierarchy then induces others who would prefer to drop out to persist. Jencks' main conclusions have since been strongly supported in the extensive studies of Hope (1984). The norm-referenced allocative (as distinct from competence-certifying) function of educational qualifications is also strongly confirmed in the previously mentioned work of Steiner (1999) and Murphy (1993). What they show is that there has in fact been little change in the competencies needed by the workforce over the years, yet everyone has to spend

more time in the so-called educational system, and accumulate more certificates, in order to get to the same place in the occupational hierarchy: The majority of US graduates now end up as maids, retail sales persons, or janitors. On the basis of his observations as a University Principal, Nuttgens (1988) suggested that one of the main functions of the system must be to "promote those who are least able to do anything" into influential positions, and McClelland (1961) showed that the system does, indeed, tend to squeeze out those high in *need* Achievement. Tomlinson and Tenhouten (1976) showed that primary schools promote a disproportionate number of those who are most willing to do whatever is necessary to secure their own advancement regardless of the ethical implications that so doing may have. They suggested that such individuals may have an important role to play in a society largely composed of a network of fraudulent systems which, like the "educational" system itself, fail to deliver what they claim to deliver and that what those concerned were actually doing was conveniently obfuscated by the educational system having denoted them as "highly able".

If such claims were true (as they probably are), one would be left with a very strong feeling that the forces Kuhn argued lay behind the hegemony of particular scientific positions (in this case the hegemony of the single-factor model of "ability") are not the only process at work here but are supported by some very strong sociological requirements that are perhaps only too apparent to those in positions of authority. Put more strongly, instead of, as Gottfredson would have us believe, promoting the most able, one of the latent functions of a single-factor model of "ability" could well be, not merely to "legitimise the rationing of privilege", but to satisfy a sociological "need" for a single and unarguable criterion of merit to legitimise a social hierarchy which contributes enormously to the network of forces which result in most people spending most of their time contributing to activities which are, directly or indirectly, destructive of other people's quality of life, and the chances of our species and the planet surviving – i.e. to activities which can only be regarded as highly unethical[6].

If that were the case, it would forcefully raise the question of how these social forces come to exert their influence.

These suggestions and this question behove us to examine the way in which multiple-talent educational programmes have been driven out of schools in a little more detail – because the bringing *in* of the "national curriculum" and its associated assessment practices has not

been the only process at work. It has been paralleled by an active move to drive multiple-talent education *out* of schools. One example was John Major's announcement that "As from tomorrow, there will be no more 'play schools'. All children will be sitting in rows facing the teacher and being taught". This drive to eliminate "open" or "progressive" education has not been limited to the state sector. It has been accompanied by a vehement campaign to undermine private schools with wider objectives, even those specifically set up to cater for those pupils who could not cope with the authoritarian monocultures of state schools. OFSTED's attack on Neill's *Summerhill* – which later turned out, as a result of an almost unique court action, to have been almost entirely fraudulent (see Stronach, 2003)–is but one of many that have, for lack of money, remained publicly unchallenged.

It is of interest that these developments followed a much earlier, but nevertheless very revealing, standardising "development". For some 15 years starting from the early 1960s, committees of the Schools' Council for Curriculum and Examinations in England and Wales debated the desirability of establishing a common system of examinations. For good reasons, they never came to a conclusion. Then a new Minister for Education established a new committee with a remit to come to a conclusion within six months. That committee (Waddell, 1978) observed that pupils had a huge variety of different talents and that these could only be fostered through very different types of educational programme. It noted that workplaces and society required a wide variety of people who possessed very different talents. It therefore (correctly) concluded that there was a need for a wide variety of different types of educational programme which would foster very different competencies and in the course of which pupils would master very different areas of knowledge. This led it to the conclusion that it would be necessary to retain a diversity of examining boards which would each promote a wide variety of courses covering different content, aiming at different goals, and assessed using different forms, or "modes", of assessment which would make it possible to give students (and thus their teachers) credit for having developed such qualities as creativity and critical thinking. Then it did an amazing thing. In one sentence embedded in a long paragraph it said "the results will be expressed on a single scale of seven points in a subject area". This, of course, negated all the proposals it had made for arrangements to promote and cater for diversity. How can the results of educational processes designed to nurture the ability to problematise, collect data,

and influence others be expressed on the same scale as the outcomes of a programme designed to inculcate the received wisdom about 18th century English history? One can only conclude that the sociological need for a single and unarguable criterion to legitimise the allocation of position and status – and with it a whole social system for rationing privilege–had somehow over-ridden educational and occupational considerations.

The examples so far given of the, largely hidden, drive to eliminate multiple-talent, competency-oriented, education are accessible in the literature. But there are many, on the surface individually amazing, examples in the (so far as I know) as yet unwritten history of teachers' attempts to bring education into schools. I have been urged to refer to more of them here. But there is a problem–and it is not just space. What happened to any individual project is largely known only to one or two people who were closely associated with it. And those people were not associated with other projects which–at least on the surface–suffered a similar fate. So, until someone systematises what happened all I can do is cite individual examples known to me … and the resulting text seems out of character with the rest of this article. Nevertheless, a few examples must be given. Some relevant projects were associated with the, largely teacher controlled, Schools Council that has already been briefly mentioned. Many of its major curriculum projects disappeared for reasons known only to a few close associates. I know at least part of the story about what happened to its Integrated Science project, which was deliberately closed because it was both encouraging pupils to think about what they were doing and ensuring that they could get credit for so doing in the examination system. I am told that a similar fate befell the Humanities Project, MACOS, and a related mathematics project. These processes were by no means limited to the UK. At much the same time, the US office of Economic Opportunity–*not* the Office of Education–initiated Headstart and Follow-Through with a view to allowing thousands of sponsors to initiate projects based on their own theories about the causes of the range of problems known to be associated with social and economic disadvantage. Some of these were enormously successful in producing change (see Raven, 1980 and 1981 for a fuller account of some aspects of this work). This presented the evaluators (e.g. Stanford Research Institute) with a problem, which they set about trying to cope with. But then an apparently extraordinary thing happened. Control of the projects was wrested from the Office of Economic Opportunity and transferred to the US Office of Education. This promptly directed

the evaluators to pay no attention to outcomes other than raising IQ, school achievement, and staying out of trouble with the police. This had the effect of forcing most of the sponsors to abandon most of their objectives. But what it is most important to note about the remaining objectives is that, while laudable, they are norm referenced and, as such, logically unobtainable by a cross-section of pupils. IQs are by definition relative to the scores of other children in the same age group. One cannot have "most" children "above average". And, as Hope (1984) also demonstrated, this particularly applies to "at risk" pupils. As soon as one moves some pupils out of "remedial" classrooms where they are "set up" to be in trouble with the police, their seats are taken by others. What one sees very clearly here is the role which the educational system, qua system (and not via the "hidden curriculum") plays in contributing directly to the cementation of a social structure that has a range of knock-on effects and the willingness of authority to intervene in, and effectively destroy, the educationally-oriented activities created by people with a genuine interest in children, people, education, development, and humane ideals in society to ensure that these sociological functions are performed.

2. Processes operating in psychology.

The second set of reasons why so little has been done to act on Spearman's insights may be that our society somehow "needs" a single and unarguable criterion of merit to operate as it does and, in particular, to enable it to progress along the self-destructive trajectory on which it has embarked. Instead of, as Gottfredson would have us believe, promoting the most able, the latent function of a single-factor model of "ability" could well be to compel all, against the threat of the destitution and subjection to the demeaning and dehumanising treatment of the so-called "welfare" services that is with so much visibility heaped on those deemed "less able", to carry out activities which, like those conducted in the educational system, are conspicuously fraudulent, unethical, and destructive of human well being and capacity to survive as a species.

The first evidence supporting this thesis to be reviewed here comes from the fact that, in the end, the McBer researchers who did most to promote recognition of the need for diversity (if not an appropriate framework to handle it) have, as I have shown in more detail in chapters in Raven and Stephenson (Raven, 2001a&b), been somehow induced to bring their framework into line with the classical paradigm. This is nowhere more striking than in the contrast between the conclusions about effective

teaching which they present in a report prepared for the Department of Education and Employment in the UK (Hay/McBer, 2000) and their earlier work on the topic (Alschuler, Tabor, & McIntyre, 1970; Alschuler, 1973; McClelland, 1982a&b), in which they dwelt on the varied, competence-based, qualities that it is important to nurture through the educational process and on the diverse ways in which teachers contribute to a system which actually educates (Huff et al., 1982). In their later work for the UK Department of Education and Employment, the McBer researchers not only accept the traditional, "single-factor", criterion for judging educational performance (which had previously been shown to be unrelated to any form of competence worth the name–see Alschuler, 1973; McClelland, 1973), they then relate teacher effectiveness, judged in terms of their ability to achieve this outcome, to what amounts to a particular presentation of the 16 competency "variables" listed in the Hay/McBer *Scaled Competency Dictionary* (Hay/McBer, 1996) using multiple regression techniques. Nothing could be more conventional. Nothing could be further removed from the kind of product which their earlier work would have led one to expect them to generate. How did this come about?

My thesis will be that this reversal was largely induced by what the so-called "market" (performing the dysfunctional functions we have noted) was willing to pay for. Some evidence supporting this claim comes from comments made by Lyle Spencer while he was at work on *Competence at Work* (L. M. Spencer & S. M. Spencer, 1993). In that book, the Spencers sought to develop a framework which would enable them to impose some kind of order on the vast range of competencies which McBer researchers had shown to be important in the course of their numerous studies of many domains of work.

According to Spencer, they set out, following suggestions made in my book *Competence in Modern Society* (Raven, 1984/1997), to develop an "atomic theory of competence". Unfortunately, the *publishers'* reviewers argued that the value and usefulness of such a framework would, because of people's prior expectations and commitments, be lost on most potential readers. This would mean that there would be little demand for the book and render its production uneconomic. And this, indeed, has been my experience with *Competence in Modern Society.* The Spencers therefore settled for the lesser task of producing a "dictionary of occupational competencies".

Further evidence that advance in academic understanding has been undermined by what will sell into current organisational structures comes

from the way in which, as I have shown in the previously mentioned chapters of *Competence in the Learning Society*, McClelland's 1958 radical measurement insights have been corrupted back into a classical "variable-based" framework. Even what remained of the original distinctive insights in *Competence at Work* has been obliterated as the contents of that book were distilled down into Hay/McBer's *Scaled Competency Dictionary* (Hay/McBer, 1996).

So far, this argument has related only to selling a conceptual framework into an academic and consulting market. But it has also proved impossible to sell the very *tools*, based on the new measurement model, which the work of Klemp, Munger, and Spencer (1997), Schön (1973, 1987), Kanter (1985), and others indicates are crucial to improved organisational performance–and the reasons are revealing indeed.

But before turning to them we may note that, for 15 years, Schön and Argyris … two of the most respected figures in occupational psychology … were unable to modify the management-development programmes at MIT to reflect the results of Schön's research (1987). The problem was not only the way in which the discipline-based, technico-rational model of competence was locked into lecturers' career structures and the assessment procedures used by the college. It also included the reactions of the students. They argued that no one could tell whether they were competent managers or not. Under such conditions, what they had to do was focus on getting themselves promoted. This, they claimed, depended on parading the latest technico-rational jargon in front of their superiors, or, in other words, doing exactly what the so-called educational system had taught them to do and selected and promoted them for doing.

To return now to the question of selling tools. Not only have we–like Taylor (1973, 1974, 1985, 1986) before us–been unable to sell our books on competence and the effective management of genuine education in commercial quantities … we have also been unable to sell the tools we developed using the theoretical framework we built up. And the reason is of more than passing interest. Despite the demonstrated importance of managers thinking about, placing, and developing the talents of subordinates (Kanter, 1985; Schön, 1983; Klemp et al., 1977; Jaques, 1976, 1989), not only do only 10-12% of British and American managers (compared with some 40% of Japanese managers) think it is important to do this, even less of them do it (Raven, 1997; Graham, Raven, & Smith, 1987). The reason these managers give for the discrepancy between their priorities and their behaviour is that they have no time to do it. But, as we have seen, those who think they should do it are a minority. *Most*

managers argue along the lines that salespeople are hired as salespeople and that they should do just that. Despite Kanter's evidence of the vital importance of the activity, they believe that salespeople should not set about telling them how to improve the product, the stock control system, and so on. That is someone else's job. They (the managers) should not have to spend time thinking about how to redeploy staff, worse, how to assemble fluid, network-based, working groups based on *part* of staff time. If they have to think about redeploying their subordinates it shows that the wrong people have been hired and should be fired.

It follows that we cannot sell the tools we have developed to help managers do their jobs without a major investment in organisational development and without corresponding change in managers' job descriptions and in the criteria adopted in the appraisal systems used to assess their competence and judge their performance.

The implications are serious. If the questions "What will sell?" and "In what kinds of courses will people be willing to enrol?" really have a major impact on the scientific paradigms and educational activities we can pursue we need to take the situation very seriously because most governments have signed General Agreements on Trade which commit them, among other things, to "privatise all services (including education) to the maximum extent possible". The effects can be expected to be nothing short of disastrous.

A re-formulation

The processes described above may be viewed as an outcome of, among other things, unthinking (or perhaps engineered) espousal of the kind of reductionist science which requires scientists to focus on establishing the strength of the relationships between one variable and one other variable at a time and to ignore all other inputs and consequences. The effect of this is to lead scientists not only to shy away from any attempt at *comprehensiveness* (claiming that it is unrealistic and "too difficult") but also to deny responsibility for examining consequences outwith those covered in the studies they have been commissioned to undertake. (The word "comprehensive" is here used to suggest an attempt to get a rough fix on *all* the effects of the experimental variation on *all* relevant outcomes.) The effect is to promote a vision of science which is both deeply unethical and lacking any form of objectivity worth the name.

Shiva (1998) has noted that the promotion of such a vision of science is somehow linked to the promotion of monocultures of mind

(in both education and in the range of scientific perspectives [theories] that are deemed acceptable), the promotion of monocultures in society, and to the promotion of monocultures in agriculture. The net result of the autopoietic[7] system constituted by these interlinked processes is the headlong plunge of our species toward its own extinction, carrying all known life with it.

Part IV: The Way Forward

A Brief Statement

At least two things would seem to follow from what has just been said. First, instead of evaluating studies primarily in terms of the accuracy of correlations established between single variables, it would seem that the main index of quality should be *comprehensiveness*. Second, it would seem that we should articulate and embrace what may be termed an ecological image of science. In this, the dominant concern would be to study and map the multiple and interacting feedback loops, intermediary outcomes in, and diverse results of, any process we seek to understand and describe. Morgan (1986) and Raven (1994, 1997) have provided partial illustrations of what such work might look like. An appropriate name for such activity can be derived from the word "cybernetics". Cybernetics is the study and design of guidance and control systems in animals and machines. It is therefore appropriate to use the term *socio-cybernetics* to refer to the study and design of guidance and control systems in society.

Socio-Cybernetics: An Illustration

I may again illustrate what I have in mind by reference to our work on the educational system.

However, by way of a preliminary comment, I must first emphasise that I do not deny the importance of many other contributory factors besides those on which I will focus. On the contrary, in contrast to what many educational philosophers would have us believe, our work has clearly shown that the effective implementation of individualised, competency-oriented, project based, enquiry-oriented education in group settings is just too difficult for most teachers without: (i) better ways of thinking about multiple talents; (ii) a better understanding of the procedures to be used to nurture those talents on a group basis; (iii)

tools to assist in the design and implementation of the individualised, competency-oriented, developmental programmes that are needed to harness pupils' individual motives and lead them to develop otherwise invisible components of competence; (iv) ways of giving pupils credit for their idiosyncratic qualities; and (v) means of giving teachers credit for their otherwise invisible educational accomplishments – and, especially, for nurturing among their pupils a wide range of diverse talents which could not possibly show up on any conventional test.

Nevertheless, our work has also revealed that many other processes are also at work. These have to do with the inability of public management systems in general to cater for diversity (Raven, 1989, 1995), their inability to release the ferment of innovation and learning that would be required to deal with the multiple causes of these over-determined problems–and especially their inability to provoke learning about the systems processes which regularly undermine well-intentioned public action, their inability to support a transformative adventure in which the outcomes cannot be specified beforehand (see also, Jackson, 1986), and their inability to initiate comprehensively evaluated experiments and change them in the light of the evaluations. The interactions between these components are mapped in Figure 19,1[8].

The Figure illustrates how the narrow nature of educational provision is heavily over-determined and demonstrates why it has been so difficult to introduce change in education. We are dealing with a *system,* or network, of hidden social forces which drive the system. The cumulative effect of these forces is that the system becomes self-perpetuating. The effects of any single change are negated by other forces and predictable reactions produced by the overall system of forces. As a result, "common-sense" reform does not work. While indicating the motives which might be harnessed to produce educational change, the Figure also shows the difficulty of linking those motives to the points at which systemic interventions might be targeted. While pervasive, *system-oriented,* change is required, so many changes are needed in every nook and cranny of the system that there is no possibility of those changes being centrally mandated.

What happens is not determined by the wishes of parents, teachers, ministers of education or anyone else but, both directly and indirectly, by the sociological functions the system performs in society. One needs to take these sociological forces seriously and ask how they can be harnessed in the way that marine engineers harness the wind: They will not go away.

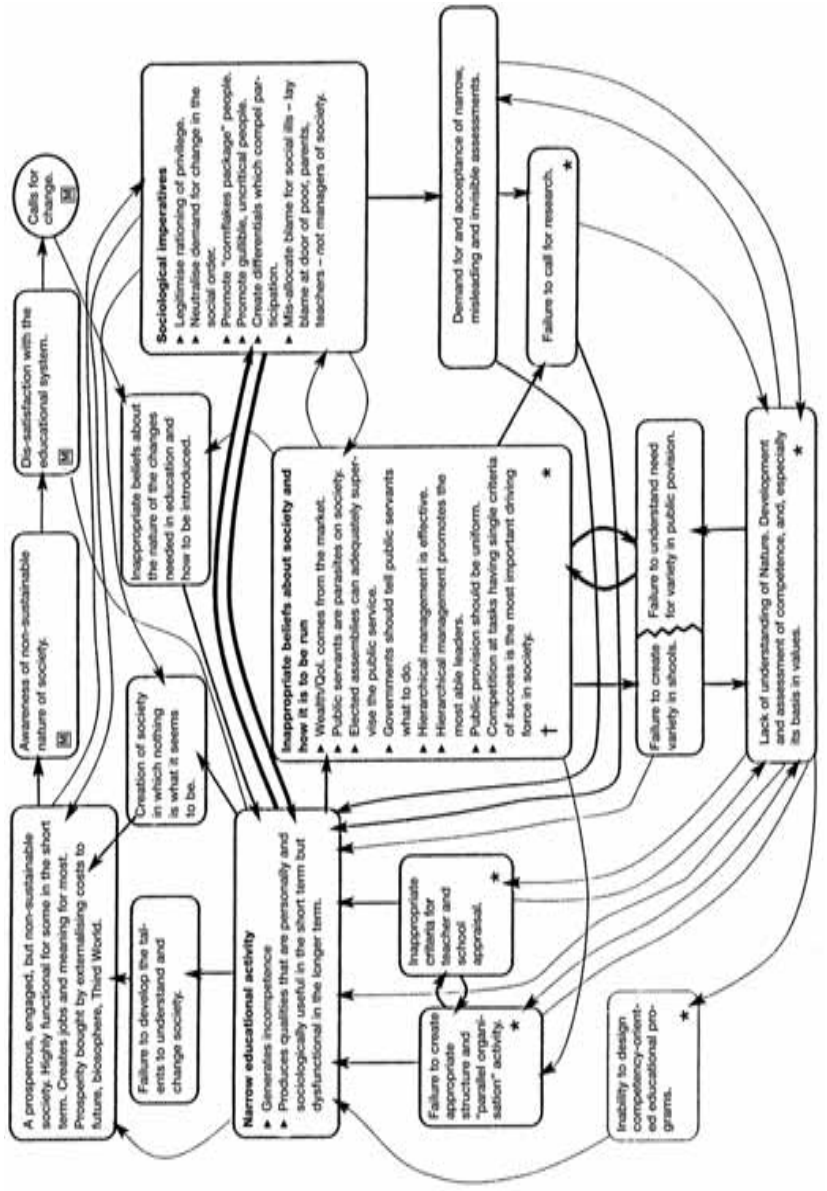


Figure 19.1 . **Feedback Loops Driving Down Quality of Education**

One effect of these forces is to *create* inappropriate beliefs about society and how it is to be managed–and these reinforce the operation of the system.

In more detail, the Figure shows:

1. That the dominance of the activities with which schools are preoccupied arises from:
   (i) A series of sociological imperatives (e.g., that schools assist in legitimising the rationing of privilege);
   (ii) Inappropriate beliefs about the nature of the changes that are needed in education itself, the management of the educational system, and the management of society;
   (iii) Society's failure to initiate research which would yield useful insights into such things as (a) the nature of competence and how it is to be fostered and (b) how to manage the educational system to nurture high-level generic competencies;
   (iv) The absence of (a) systematically generated variety in, and choice between, educational programmes which have demonstrably different consequences and (b) Information on the consequences of each of these alternatives;
   (v) Failure to introduce "parallel organisation activity" to produce innovation within schools, and
   (vi) Inadequate dissemination of the results of research into the nature, development, and assessment of generic high-level competencies, and, especially, the implications of the values basis of competence.
2. That this narrow educational process has a series of knock-on effects which finally contribute to its own perpetuation. The competencies and beliefs that are nurtured and inculcated in schools reinforce a social order which offers major benefits to "able" people who do what is required of them without questioning that order; it creates endless work which gives meaning to people's lives (but does not enhance the general quality of life); it creates wealth at the expense of the biosphere, future generations, and the third world; and it protects its citizens from a knowledge of the basis of their wealth. The educational system helps to teach a host of incorrect beliefs which collectively result in nothing being what it is popularly or authoritatively said to be (for example, the educational system itself claims to be about promoting the growth of competence when it in fact mainly operates to engage vast numbers of people in "teaching" and "learning" activities of little

educational merit but which ensure that those who are most able and willing to challenge the fraudulent nature of the system are routed to social positions from which they can have little influence while those who are least able to do anything except secure their own advantage are promoted into influential positions in society). This double-talk makes it extremely difficult to conduct any rational discussion of the changes needed in society. The sociological imperative that schools help to legitimise the rationing of privilege helps to create a demand for, and encourages acceptance of, narrow, invisible, and mislabelled assessments. Those predisposed to acquire these "qualifications" are not inclined to see the need for, or to commission, genuine enquiry-oriented research or notice other talents in their fellows. Teachers who become aware of the hidden competencies of their "less able" students experience acute distress. The lack of understanding of the nature of competence leads to a failure to underline the need for a variety of value-based educational programmes and thus to the perpetuation of narrow educational activity.

3. That the main motives for change are widespread awareness that there is something seriously wrong with the educational system, and, more specifically, that it fails miserably in its manifest task of identifying, nurturing, recognising, and utilising most people's motives and talents. The most commonly proposed solutions to this problem, based as they are on other misunderstandings, are, however, inappropriate. Another motive for change stems from increasing recognition that we have created a non-sustainable society and that basic change in the way society is run is essential.
4. That there are a number of points at which it should be possible to intervene in the feedback loops to create an upward spiral. These might involve:
   (i) Promoting wider recognition that one cannot get value for human effort in modern society unless we introduce better means of monitoring and evaluating the long-term effects of what we are doing and better ways of giving effect to information on such effects. This points to the need to change the way we run society, to the need to introduce more, and more appropriate, social research and evaluation activity, and to find ways of holding public servants and politicians accountable for seeking out and acting on information in an innovative way in the long-term public interest;

(ii) Introducing the "parallel organisation" activities that are required to promote innovation within schools;
(iii) Establishing a greater variety of distinctively different, value-based, educational programmes and providing information on the short and long-term, personal and social, consequences of each;
(iv) Creating public debate about the forms of supervision–the nature of the democracy–needed to ensure that public servants seek out and act on information in an innovative way in the public interest and,
(iv) Disseminating what is already known about the nature, development, and assessment of competence and its implications.

Implications for the Role of the Psychologist

In developing this map, we have attempted to follow the injunctions of House (1991), Parlett (1972, 1976), and Hamilton et al. (1977) to use psychological data to illuminate the hidden network of social forces which overwhelmingly determines our behaviour and our theories. Many readers will claim that, as psychologists, we should not have done this or that we have "gone way beyond our data" in doing it. Yet, if we, as psychologists, wish to claim either to be serious students of the determinants of behaviour or that we aspire to the application of our science to benefit society, there is no doubt that we need to take the study of such forces seriously[9]. They do, indeed, strongly determine human behaviour, they are to be illuminated by using psychological data in appropriate ways, and the only way to intervene in them is by adapting the results of psychological research into effective organisational arrangements and human competence and using it to develop new organisational arrangements and information-based management tools.

But we will not engage with this task if we continue to work within the constraints and shared images of the role definition that we have accepted in the past. We need to actively articulate and promote a new role for ourselves.

The Wider Context: The Destruction of Life on Earth

There is not space in this article to develop in any detail the claim that the autopoietic system we have mapped for the educational system is part of a wider autopoietic system that is heading our species toward its extermination carrying all known life with it.

Yet it is now widely recognised that we, as a species, are heading toward our own extinction (Oskamp, 2000; Stern, 2000; Raven, 2001c; Anderson, Douglas, Holmes, Lawton, Walker, & Webb, 2001). Although Oskamp cites numerous trends that are accelerating exponentially out of control, the most striking is Wackernagel and Rees (1996) demonstration that it would require five back up planets engaged in nothing but agriculture for everyone alive in the world today to live as Americans do.

There is a strong tendency to attribute this plunge of homo-sapiens toward self-destruction, despite widespread recognition of the need to radically change the way we live, to the doings of evil capitalists. Yet our work on the educational system shows that the process has too many components to support the view that it has been designed by an evil elite. What is most striking is that the system has evolved further and further along its current trajectory despite the repeated demonstration that the vast majority of pupils, parents, teachers, ex-pupils, and employers want it to move in exactly the opposite direction.

This claim parallels that offered by Galbraith (1992) in his quest for an explanation of the great financial crash of 1929. A search for evil people on whom to pin the blame gets us nowhere. What one sees is in the great crash is the cancerous growth of an emergent autopoietic system which no one can see how to stop until the system as a whole collapses. Morgan (1986) has developed a socio-cybernetic diagram for inflation more generally … and shown that there *are* a number of negative feedback loops which could be amplified to damp the system down.

I have elsewhere (Raven, 1997; Raven & Stephenson, 2001) developed a socio-cybernetic map of the processes that are driving our whole society, against our will, toward our self destruction, but to introduce it here would be to raise questions which would take us far beyond the scope of this paper.

The Way Forward: An Illustration by Analogy

In considering what needs to be done to get out of the messes we are in it may be helpful to pursue an analogy from physics.

Prior to Newton, if things moved it was because they were possessed of animal spirits … they were *animated.* Likewise, prior to Newton, it was impossible for sailing boats to sail into the wind. Newton made three crucial observations: (1) If things moved (or changed direction, or stopped moving) it was because they were pushed or pulled; (2) To every force there is always an equal and opposite reaction; the problem is to identify it, and (3) The forces acting upon a body can be resolved into orthogonal components.

The first of these implied that the wind was not animated. Instead of praying to the gods for a favourable wind, one should set about trying to harness the forces which, up to that point, had simply crashed boats against the rocks to do useful work.

The second observation implied that there must be somewhere an equal and opposite force to the wind. A quest to identify that force led to its being found–unimaginably–in the sea. And a search for ways of harnessing that force led to the addition of keels to sailing boats.

The third observation led to the realisation that the opposing reactions of the wind and the sea could be resolved into a component pushing, if not directly into the wind, at least in a direction which enabled one to tack into it.

These remarks imply that the first thing we have to do is to de-animate the forces that are seen as driving us toward our self destruction. We have to stop blaming (and wringing our hands about) our leaders and the capitalists[10]. Instead, we have to see them as *expressions* of a network of hidden forces. They are selected and promoted and behave as they do because of those forces. What is more, people who behave in ways which resemble our leaders and capitalists are not few in number but pervade our society. Then we have to identify those forces. And, after that, take steps to harness them. A relatively naïve suggestion (which nevertheless illustrates the point) is that including measures of a wider range of the outcomes of education in the certification and placement processes used by schools would drive schools towards doing the things we want them to do rather than away from them. (Such a development would be the equivalent of adding keels to sailing boats.)

But the development of a relatively safe network of sailing boats depended on many other things besides the classic academic inputs of

Newton and others. It also depended on the emergence of a complex socio-cybernetic system: It was necessary to accumulate a host of charts of the seas and the ports, to evolve sextants and chronometers so that ships' captains could know where they were on the high seas, to erect lighthouses, to develop means of paying lighthouse keepers, and so on and so on.

Parts of this system evolved relatively naturally, but other parts–such as the development of chronometers–required enormous purposeful public investment.

There is one more point to be drawn out of this analogy. Many have asked "Are we strong enough to fight these dominators; these capitalists and politicians?"

This is analogous to asking "Are we strong enough to fight the wind?" It is the wrong question. What we have to do is to understand and to map the relevant socio-cybernetic systems and then use our insights to develop alternatives. As numerous scientists have discovered over the course of history, the personal costs of challenging conventional authority can be enormous. But collectively–and with superhuman individual contributions–it was accomplished. To us now falls the mantle of carrying the process forward. We, as psychologists, need to set about bringing into being the kind of paradigm shift that was brought about by Newton and his colleagues. It demands classic academic activity. But we also need to initiate and contribute to the wider developments that are required to evolve a more appropriate socio-cybernetic system to manage society.

Mapping socio-cybernetic feedback loops has proved a daunting task. Despite the work of Morgan (1986), improving on Figure 19.1, depicting the forces that are contributing the elimination of the species and the destruction of the planet, and clarifying how to move forward has proved difficult indeed (see Raven & Navrotsky, 2000). Certainly it has not proved as easy as either Morgan or Navrotsky suggested to identify the negative feedback loops that damp down the operation of the system with a view to amplifying them in order to bring about desired changes.

Developing a specification for an alternative socio-cybernetic system for the management of society is a still more daunting task. When discussing the results of our attempt to map the interlocking network of feedback loops that perpetuate our dysfunctional educational system I mentioned, although I did not elaborate the point, that, if we are to move forward, we need to design a better public management system for society, that is to say, to design new forms of public management that will

operate in the long-term interest of the general public instead of in the short-term interests of dominators.

The requirements to be met by such a design can be found in the work of Adam Smith and Fred Hayek. One of the key observations they made was that, contrary to what almost everyone believes and assumes, the system should work *without* leaders deemed to be wise. The reason was simple, but devastating in its implications: *There can be no such thing as a wise man or woman.* The reason was again simple: the key knowledge required to take informed decisions – knowledge of what will happen as a myriad of current developments come together–cannot be available to anyone. Stated in one way this means that the system has to work without assuming that some person or group of people can know anything very much. As Smith and Mill repeatedly asserted, government decision-taking cannot be other than decisions by committees of ignoramuses. Put another way, the design specification is that the system must harness the expert information that lies in the heads, hearts, and hands of billions of people – hearts and hands because much of the information is not verbalised and consists of feelings and unverbalised knowledge of how to do things. In yet other words, an acceptable design must be non-authoritarian and make provision for widespread experimentation accompanied by many interacting feedback (learning) loops. There is not space here to show that our current societal management system – although often described as a market management system – actually operates in entirely the opposite way; that we live in a managed economy in which the function of money has been reversed. Instead of providing the basis for an "invisible" management system in which billions of people vote with their pennies on a myriad of issues, the control of cash flows and the determination of prices is used to achieve goals determined by the trans-national corporations and through the politico-bureaucratic process. Nor is this the place to show why Smith and Hayek's "market" solution to the design problem they correctly identified does not and cannot work (Raven, 1995). And neither is this the place to outline the new arrangements that are required in any detail (a fairly detailed sketch can be found in Raven, 1995). But at the heart of the requisite new arrangements–this designed socio-cybernetic system—seem to lie new concepts and forms of bureaucracy and democracy … new organisational arrangements about which psychologists (following the work of Kanter, 1985 and Schön, 1973, 1983, 1987) should have much to say. And new job descriptions and appraisal systems for public servants. In other words,

the development of a new socio-cybernetic system for the management of society depends centrally on the application of the concepts and methods of organisational psychology to the running of society.

But there is one more, somewhat paradoxical, thing to be said. Proceeding as we have suggested here essentially involves turning psychology inside out. It means de-animating human behaviour and, in a sense, attributing behaviour to the hidden social forces that act upon us. Of course that is an over-statement because we have spoken of the role of these forces in selecting and promoting certain sorts of people. Nevertheless there is something of an irony in suggesting that the way forward involves promoting the use of psychology to depsychologise human behaviour.

Summary and Conclusions

In the course of this paper we have seen that other qualities besides ***g***, and especially the ability to contribute in one currently invisible way or another to group processes and the ability to understand and intervene in the external, social-systems, determinants of behaviour, *are* vitally important.

The invisibility of such contributions is produced in part by a network of interacting, but mutually supportive, processes which include deficiencies in our traditional psychometric paradigm and the procedures used to assess performance and also from assumptions about the efficiency of hierarchy. But, most importantly, it stems from the sociological need for a single and unarguable criterion of merit to legitimise a social hierarchy which contributes enormously to the network of forces which result in most people spending most of their time contributing to activities which are, directly or indirectly, destructive of other people's quality of life, and the chances of our species and the planet surviving–i.e. to activities which can only be regarded as highly unethical.

This observation prompts two more basic conclusions: (i) The *main* determinants of behaviour are external rather than internal; if psychologists wish to continue to claim special competence in relation to understanding behaviour, it will behove them to pay attention to these external forces, and (ii) If psychologists wish to understand these social forces, and, in particular, to assist in the development of arrangements which will enable society to achieve its goals more effectively, then it is

essential to find ways of illuminating the socio-cybernetic systems that control the operation of society and to use this information to generate designs for more effective arrangements for the management of society.

More specifically, it would seem that the relationships so strikingly portrayed in Gottfredson's paper have come about, not because they contribute to a getting useful work done in an effective way for the benefit of society, but for precisely the opposite reason: They contribute to a network of myths, thoughtways, hidden social forces, and actions which obscure and render invisible the processes actually at work–and those processes result in such destruction of the planet that its very survival is in jeopardy.

From a practical point of view, the article highlights the need for better tools to help parents, teachers, and managers think about, develop, and utilise the vast array of talents that people possess and that are needed in society. It points to the need for a better understanding of the nature of developmental environments and the tools that are needed to organise them. But, most importantly, it underlines the need to develop more appropriate organisational arrangements, job descriptions, and organisational and staff appraisal systems for the various domains of policy–such as the educational system–that are required to run society in the long-term public interest. The development of these specifications and tools is quintessentially a task for psychologists[11].

Notes

1. Lester (2001) has delightfully argued that, despite almost universal acceptance of the contrary belief, *knowledge* is the one thing that that one *cannot* assess because it is largely idiosyncratic and tacit. Gottfredson (2003) has also drawn attention to the nonsense of Sternberg trying to assess "tacit" knowledge using tests of explicit knowledge.

2. As discussed more fully in Raven (1991) there is no sense in which a traditional "science" test assesses the competencies of the scientist: The ability to problematise, conceptualise, locate relevant earlier work, familiarise oneself with the relevant theories, built up a unique store of up-to-date specialist information, invent ways of collecting data, gain help, raise funds, find ways of summarising data, and so on. Nor does it provide an index of knowledge of any kind of genuine sample of scientific facts … since, with the knowledge explosion, this domain is vast. Instead,

performance on these tests reflects only the ability to retain for a short while, and regurgitate, a sample of facts chosen by an authority (i.e., abdication of responsibility for learning and evaluation of the relevance of what is learned and/or the ability to present the material in a way which will appeal to the examiner–i.e., a concern with self-presentation rather than scientific knowledge). Likewise there is no sense in which the ability to answer nine questions about a paragraph unconnected to the respondent's knowledge will index any meaningful reading competence since that depends on such things as the ability to find information related to one's purposes, to use that information to provoke lateral thinking, to evaluate and escape from blind alleys–that is to say to refuse to read and try to understand the irrelevant. It follows that, for these and related reasons, most of the tests in common use cannot be said to measure that which they purport to measure. Bluntly, they lack construct validity.

3. It is often asserted (e.g. by Weiner, 1992 and Snow et al., 1996) that these measures have been discredited. However, when one looks at the studies that are cited, one discovers that the operational definitions of the relevant constructs are entirely different to those deployed in the studies conducted by McClelland and his colleagues. One set, for example, uses the "Achievement Motivation" Scale of the Edwards Personal Preference Schedule. This is a Likert-type scale which essentially asks people how much they are attracted to a number of activities which might be viewed as being "achievement oriented". In no sense does it assess whether or not people are likely to bring to bear the cumulative and substitutable competencies that are required to carry out achievement-oriented activity effectively. Many of the measures used in the so-called validation studies are even more reductionist, consisting of such things as single (not even multiple) value-expectancy measures.

4. It may be useful to underline the full significance of this observation. What it means is that those studies that are widely used to support "evidence based policy" in education and "evidence based treatment" in drugs-based health care, psychotherapy, criminology, and agriculture in reality contribute precisely the opposite. They are the least scientific, objective, and ethical studies imaginable. By failing to report numerous, often disastrous, personal and social, short and long term, consequences of the policies and programmes being legitimised they contribute directly to the implementation of the unsupportable. And they do so under the guise of scientific respectability. This is why Shiva (1998) has argued that reductionist science contributes directly to monocultures … not only in social culture … but also in agriculture and in mind itself. Nothing could better illustrate the way in which various social processes, including the use of words to conjure up images that are precisely the contrary of what is actually refered to, combine to head us in a direction in which no rational person would choose to go.

5. Although only one or two good studies are cited in each case, many more are reviewed in Raven (1994).

6. *Most* work in modern society is highly unethical. As spelt out in Raven (1995) it involves doing such things as:

   - contributing taxes, research, or direct manufacturing activity to a war machine which not only directly takes the lives of hundreds of thousands of people each year but also consumes and/or destroys huge quantities of planetary resources in manufacturing or training exercises or as a result of dumping "waste products" arising from the manufacture or usage of nuclear and other weapons;
   - producing, marketing, or distributing junk foods, junk toys, and junk cars. The manufacture of these unnecessary commodities consumes enormous quantities of irreplaceable resources and generates waste which cannot be effectively disposed of. It therefore contributes enormously to the destruction of the soils, seas and atmosphere. Distributing them involves flying almost identical goods in opposite directions all over the planet and centralised distribution arrangements which depend on trucks, cars, and the construction of highways which also generate enormous pollution. Production also results in massive exploitation of labour and not only in "third world" but also at home. Marketing produces needs which cannot be satisfied and thus leads to debt and dis-satsifaction among huge sectors of the population;
   - offering junk education and junk research. Junk education fails to develop, and, as shown in this article, renders invisible, most people's talents thereby denying them an opportunity to develop and use them. The neglected talents are those that are most important from the point of view of reforming our way of life so that the species and the planet have a chance of survival. The system also generates feelings of inadequacy in vast numbers of people and labels them as "unemployable", suitable only for degrading and dehumanising treatment by the so-called "welfare" services. Junk research occupies the time of millions of people–and not only those directly involved in the research or in reviewing grant applications and the resulting publications, but also in building and maintaining the "necessary" buildings, printing presses etc.;
   - contributing to a drugs-based health care system that destroys all caring worth the name and diverts attention away from the societal reforms that are really necessary;
   - contributing to banking and insurance systems which are organised in such a way as to have the maximal effect from the point of view of sucking resources from the third world and exploiting–that is, destroying the lives and livelihoods of–billions of people and also reducing vast numbers of people in our own society to destitution, deprived of adequate communal care;

- contributing to energy-intensive chemicals-based agriculture whose effect is to destroy the soils, the seas, and the atmosphere as well as allocating billions of people to lives of degradation, humiliation and starvation.

In passing, it is important to note that those in the WTO and elsewhere who push through single-factor oriented educational reforms very clearly see the need to have a mythology and a social process which compels so many people to do so many things that they know to be wrong and, indeed, not even in their own best interests because the activities in which they are engaged destroy their own quality of life.

7. Autopoietic: from autopoiesis: A process whereby a system constitutes and maintains its own organisation.

8. Figure 1 merits detailed discussion which has had to be omitted here but can be found in Raven (1994, 1995).

9. In this context it may be helpful to note that, although once ridiculed for having made unjustifiable leaps of logic, geologists these days would have no hesitation in inferring from the existence of such apparently disparate things as terminal moraines and hanging valleys that the area in which they occur must once have been glaciated.

10. Readers might be forgiven for imagining that this would have been the central task of sociology. Unfortunately, just as many psychologists have been blinded by naïve theorising, so most sociologists have been prevented from engaging in any serious enquiry by a bastardised form of Marxism. The collapse of the Eastern bloc is widely–if incorrectly–thought to discredit Marxist analysis in general and not just the bastardised version of Marxism that has been mentioned. Unfortunately, this collapse has led to the abandonment of even those feeble attempts that existed within the field to clarify and map the processes we are concerned with here.

11. I have long argued that we need to move toward what might be described as more ecological ways of thinking about human behaviour: Where would biologists have got to if they had sought to summarise all the variation between animals in terms of 1, 2, or 16 variables, the environments in terms of 10, and then study the interactions between them using multiple regression techniques? But, in reality, biologists have had enormous difficulties fending off every bit as reductionist approaches as those employed in psychology. This is nowhere more apparent than in simplistic assertions about evolution by natural selection and the determination of physical structures by genes. Biologists like Waddington (1969, 1975) have had a hard time of it. Not only are bodily structures determined by the interacting effects of multiple genes (and not single ones) they are also determined by what actually emerges in a particular environmental context in locations both proximal to and distant from any particular physical location at a particular point in time. Clearly, we need some continuously interacting

model of this sort to think about human development more generally so that we cease to assert that a causes b. And we need to generalise the same model even more generally when we come to think about the development of a society composed of multiple niches. In this context, my attention has been drawn to the work of Fischer (1998) and Thelen and Smith (1998). At this point in time, I can only say that, if biologists have had a problem fending off reductionist evolutionary theories and understandings of genetics, the task of generating procedures which will enable us to think about developments in society is even more daunting.

References

Adams, E., & Burgess, T. (1989). *Teachers' Own Records*. Windsor, England: NFER-Nelson.

Alschuler, A. S. (1973). *Developing Achievement Motivation in Adolescents*. Englewood Cliffs, NJ: Educational Technology Publications.

Alschuler, A. S., Tabor, D., & McIntyre, J. (1970). *Teaching Achievement Motivation*. Middletown, CT: Education Ventures Inc.

Anderson, A, Douglas, K., Holmes, B., Lawton, G., Walker, G., & Webb, J. (Eds.) (2001, April 28). Judgement day: There are only angels and devils. *New Scientist*, Global Environment Supplement.

Bachman, J. G., O'Malley, P. M., & Johnston, J. (1978). *Adolescence to Adulthood: Change and Stability in the Lives of Young Men*. Ann Arbor, MI: Institute for Social Research.

Black, P. (1998). Learning, league tables and national assessment: Opportunity lost or home deferred? *Oxford Review of Education, 24,* 57-68.

Desjardins, C., & Huff, S. (2001). On the leading edge: Competencies of outstanding community college presidents (Chapter 8). In J. Raven & Stephenson, (Eds.), *Competence in the Learning Society.* New York: Peter Lang.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109-132.

Fischer, K. (1998) Dynamic systems theories. In W. Damon, *Handbook of Child Psychology, Vol. 1.* Chichester, NY: Wiley.

Flanagan, J. C. (1978). *Perspectives on Improving Education from a Study of 10,000 30-Year-Olds*. New York: Praeger Publishers.

Galbraith, J. K. (1992). *The Great Crash 1929*. London: Penguin Books.

Gardner, H. (1987). Developing the spectrum of human intelligence. *Harvard Education Review, 57*, 187–193.

Gardner, H. (1991). *Intelligence in Seven Phases*. Harvard Project Zero.

Goodlad, J. (1983). *A Place Called School*. New York: McGraw Hill.

Goodman, P. (1962). *Compulsory Mis-Education*. London: Penguin Books.

Gottfredson, L. S. (1997). Why ***g*** matters: The complexity of everyday life. *Intelligence, 24*, 79–132.

Gottfredson, L. S. (2003). Dissecting practical intelligence theory: Its claims and evidence. *Intelligence, 31(4)*, 343-397.

Graham, M. A., Raven, J., & Smith, P. C. (1987). *Identification of high-level competence: Cross-cultural analysis between British, American, Asian and Polynesian labourers*. Unpublished manuscript: BYU Hawaii Campus: Department of Organizational Behavior.

Grannis, J. C. (1983). Ecological observation of experimental education settings. *Environment and Behavior, 15*, 21–52.

Hamilton, D., Jenkins, D., King, C., MacDonald, B., & Parlett, M. (Eds.). (1977). *Beyond the Numbers Game*. London: MacMillan Education.

Hatch, T. C., & Gardner, H. (1990). If Binet had looked beyond the classroom: The assessment of multiple intelligences. *International Journal of Educational Research, 4,* 15–429.

Hay/McBer (1996). *Scaled Competency Dictionary, 1996.* Boston: Hay Group.

Hay/McBer (2000). *Research into Teacher Effectiveness. Phase II Report: A Model of Teacher Effectiveness*. London: Department of Education and Employment.

Hope, K. (1984). *As Others See Us: Schooling and Social Mobility in Scotland and the United States*. New York: Cambridge University Press.

House, E. R. (1991). Realism in research. *Educational Researcher, 20*, 2–9.

Huff, S., Lake, D., & Schaalman, M. L. (1982). *Principal Differences: Excellence in School Leadership and Management*. Boston: McBer and Co.

Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology; Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124(2)*, 262-274.

Jackson, P. W. (1986). *The Practice of Teaching*. New York: Teachers College Press.

Jaques, E. (1976). *A General Theory of Bureaucracy*. London: Heinemann.

Jaques, E. (1989). *Requisite Organization*. Arlington, VA: Cason Hall and Co.

Jencks, C., Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., Heyns, B., & Michelson, S. (1973). *Inequality: A Reassessment of the Effect of Family and Schooling in America*. New York: Basic Books; London, England: Penguin Books.

Jensen, A. R. (1998). *The **g** Factor: The Science of Mental Ability.* Westport, CN: Praeger.

Joint Committee on Standards for Educational Evaluation (1981). *Standards for Evaluations of Educational Programs, Projects and Materials*. New York: McGraw Hill Book Co.

Kanter, R. M. (1985). *The Change Masters: Corporate Entrepreneurs at Work*. Hemel Hempstead: Unwin Paperbacks.

Klemp, G. O., Munger, M. T., & Spencer, L. M. (1977). *An Analysis of Leadership and Management Competencies of Commissioned and Non-Commissioned Naval Officers in the Pacific and Atlantic Fleets*. Boston: McBer.

Krechevsky, M., & Gardner, H. (1990). Multiple Intelligences, multiple chances. In D. Inbar (Ed.), *Second Chance in Education: An Interdisciplinary and International Perspective* (pp. 69-88). London: Falmer Press.

Lester, S. (2001). Assessing the self-managing learner: A contradiction in terms (Chapter 26). In J. Raven & J. Stephenson (Eds.), *Competence in the Learning Society.* New York: Peter Lang.

McClelland, D. C. (1973). Testing for competence rather than for "intelligence". *American Psychologist, 28*, 1–14.

McClelland, D. C. (1982a). What behavioral scientists have learned about how children acquire values. In D. C. McClelland (Ed.), *The Development of Social Maturity*. New York: Irvington Press.

McClelland, D. C. (1982b). *Education for Values*. New York: Irvington.

McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1958). A scoring manual for the achievement motive (Chapter 12). In J. W. Atkinson (Ed.), *Motives in Fantasy, Action and Society*. New York: Van Nostrand.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50(9)*, 741-749.

Morgan, G. (1986). *Images of Organization*. Beverly Hills, CA: Sage.

Murphy, J. (1993). A degree of waste. *Oxford Review of Education. 19,* 9-31.

Nuttgens, P. (1988). *What Should We Teach and How Should We Teach It?* Aldershot: Wildwood House.

Oskamp, S. (2000). A sustainable future for humanity? *American Psychologist, 55(5)*, 496–508.

Parlett, M. (1972). Evaluating innovations in teaching. In H. J. Butcher and E. Rudd (Eds.), *Contemporary Problems in Research in Higher Education*. New York: McGraw Hill.

Parlett, M. (1976). Assessment in its context. *Bulletin of Educational Research*: *Evaluation and Assessment*, *11*, Summer.

Raven, J. (1980). *Parents, Teachers and Children*: *An Evaluation of an Educational Home Visiting Programme*. Edinburgh: Scottish Council for Research in Education. Distributed in North America by the Ontario Institute for Studies in Education, Toronto.

Raven, J. (1981). Early intervention: A selective review of the literature. *Collected Original Resources in Education, 5*, F1C6.

Raven, J. (1989). Democracy, bureaucracy and the psychologist. *The Psychologist, 2(11)*, November, 458–466.

Raven, J. (1991). *The Tragic Illusion*: *Educational Testing*. New York: Trillium Press.

Raven, J. (1994). *Managing Education for Effective Schooling*: *The Most Important Problem Is to Come to Terms with Values*. Unionville, New York: Trillium Press.

Raven, J. (1995). *The New Wealth of Nations*: *A New Enquiry into the Nature and Origins of the Wealth of Nations and the Societal Learning Arrangements Needed for a Sustainable Society*. Unionville, New York: Royal Fireworks Press; Sudbury, Suffolk: Bloomfield Books.

Raven, J. (1984/1997). *Competence in Modern Society*: *Its Identification, Development and Release*. Unionville, New York: Royal Fireworks Press (1997); First published in London, England: H.K.Lewis (1984).

Raven, J. (2001a). The McBer Competency Framework (Chapter 9). In J. Raven & J. Stephenson (Eds.), *Competence in the Learning Society.* New York: Peter Lang.

Raven, J. (2001b). The McClelland/McBer Competency Models (Chapter 15). In J. Raven & J. Stephenson (Eds.), *Competence in the Learning Society.* New York: Peter Lang.

Raven, J. (2001c). Psychologists and sustainability. *American Psychologist,* 56, 455–457.

Raven, J., Johnstone, J., & Varley, T. (1985*).* Opening the Primary Classroom. Edinburgh: Scottish Council for Research in Education.

Raven, J., & Navrotsky, V. (2000). *The Development and Use of Maps of Socio-Cybernetic Systems to Improve Educational and Social Policy, with Particular Reference to Sustainability.* Paper presented to a meeting of Research Committee No. 51 of the International Sociological Association, Panticosa, Spain, June 2000.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 4*: *The Advanced Progressive Matrices*. Oxford, England: Oxford Psychologists Press; San Antonio, TX: The Psychological Corporation.

Raven, J., & Stephenson, J. (Eds.). (2001). *Competence in the Learning Society.* New York: Peter Lang.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than ***g***. *Journal of Applied Psychology, 79*, 518–524.

Russ-Eft, D., & Brennan, K. (2001). Leadership competencies: A study of leaders at every level in an organization (Chapter 7). In J. Raven & J. Stephenson (Eds.), *Competence in the Learning Society.* New York: Peter Lang.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124(2)*, 262–274.

Schön, D. (1973). *Beyond the Stable State*. London: Penguin.

Schön, D. (1983). *The Reflective Practitioner*. New York: Basic Books.

Schön, D. (1987). *Educating the Reflective Practitioner*. San Francisco, CA: Jossey-Bass.

Shiva, V. (1998). *Biopiracy*: *The Plunder of Nature and Knowledge.* London: Green Books.

Snow, R. E., Corno, L., & Jackson, D. (1996). Individual differences in affective and conative functions. In D. C. Berliner, & R. C. Calfee, *Handbook of Educational Psychology* (pp.243-310). York: MacMillian/Prentice Hall.

Spearman, C. (1927). *The Nature of "Intelligence" and the Principles of Cognition* (Second Edition). London, England: MacMillan.

Spencer, E. (1979). *Folio Assessments or External Examinations?* Edinburgh: Scottish Secondary Schools Examinations Board.

Spencer, L. M., & Spencer, S. M. (1993). *Competence at Work*. New York: Wiley.

Steiner, D. (1999). Searching for educational coherence in a democratic state. In S. L. Elkin & K. E. Soltan (Eds.), *Citizen Competence and Democratic Institutions.* University Park, PA: Pennsylvania State University Press.

Stern, P. C. (2000). Psychology and the science of human-environment interactions. *American Psychologist, 55(5)*, 523-530.

Stronach, I. (2003). Summerhill School versus Ofsted: An update. *Research Intelligence, 82*, February, 29-30.

Taylor, C. W. (1973). Developing effectively functioning people–the accountable goal of Multiple Talent Teaching. *Education, 94(2)*, November/December, 99–110.

Taylor, C. W. (1974). Multiple talent teaching. *Today's Education*, March/April, 71–74.

Taylor, C. W. (1985). Cultivating multiple creative talents in students. *Journal for the Educationally Gifted, VIII(3)*, 187–198.

Taylor, C. W. (1986). Cultivating simultaneous student growth in both multiple creative talents and knowledge. In J. S. Renzulli, *Systems and Models for Developing Programs for the Gifted and Talented*. Connecticut: Creative Learning Press.

Thelen, E., & Smith, L. B. (1998) Dynamic systems theories. In W. Damon, *Handbook of Child Psychology, Vol. 1.* (chapter 10, pp. 563-634). Chichester, NY: Wiley.

Tomlinson, T. M., & Tenhouten, D. (1976). *Awareness, Achievement Strategies and Ascribed Status of Elites*. Washington, DC: National Institute of Education. Unpublished Report.

Wackernagel, M., & Rees, W. E. (1996). *Our Ecological Footprint*: *Reducing Human Impact on the Earth*. Philadelphia: New Society Publishers.

Waddell, J. (1978). (Chairman). *School Examinations*. London: HMSO.

Waddington, C.H. (1975). *The Evolution of an Evolutionist.* Edinburgh: Edinburgh University Press.

Waddington, C.H. (1969). *Towards a Theoretical Biology.* (2 vols.) Edinburgh: Edinburgh University Press.

Weiner, B. (1992). *Human Motivation*: *Metaphors, Theories and Research.* Newbury Park, CA: Sage.

Willis, P. (1977). *Learning to Labour*. Farnborough: Saxon House.

Chapter 20

Psychometrics, Cognitive Ability, and Occupational Performance*

John Raven

Overview

In two previous articles (Raven, 1989b, 2000), I reviewed studies suggesting that, contrary to what Flynn (1987) would have us believe, the *Raven Progressive Matrices* measures psychological abilities of fundamental importance, and that the steadiness in the improvement in these abilities over time and the similarity in the norms obtained in many – but not all – cultures at any point in time reinforce this conclusion.

In this article I will summarise remarkable new evidence that the *Raven Progressive Matrices* is measuring an important aspect of cognitive functioning. Thereafter, I will return to the question of the extent to which it measures "intelligence" (and competence more generally). This will lead to a re-examination of the test's construct validity. This discussion has important practical implications because it underlines the need to situate eductive ability scores in the context of a yet-to-be-developed framework for thinking about the wider aspects of intelligence and competence. At the same time, it raises serious questions about the way we think about the procedures to be used to establish the validity of a test and the ethics of insufficiently *comprehensive* assessment – however *in*valid some of the necessary assessments may be. The article concludes by outlining some of the parameters which must be satisfied in seeking to develop a better framework for thinking about competence and its assessment.

***

Raven (1989b) argued that the reproducibility of the psychometric properties of the RPM across different socio-economic and ethnic groups,

* Earlier versions of this article were published in S. M. Wechsler, & F. S. R. Guzzo, (1999) Psychological Assessment: International Perspectives, Sao Paulo, Brazil: Casa do Psicologo, and in the (Croatian) Review of Psychology, Vol. 7, (2000).

the regularity in the increase in scores over time, and the similarity in the norms obtained in many different cultures at any point in time all suggested that the RPM measures something of fundamental psychological importance. This theme was developed further in Raven (2000) and Raven, Raven and Court (1998, updated 2003; 2000, updated 2004), where the increases over time and the new tests developed to restore their discriminative power are discussed more fully. In this section, I will muster evidence suggesting that the RPM directly taps one important aspect of psychological functioning and that this is what most psychologists refer to as "cognitive functioning". Later I will argue that, although this is indeed the case, this label misleads – for what is generally regarded as "cognitive" functioning is primarily affective and conative. It is therefore more appropriate to claim that the RPM measures "eductive" ability – at least in relation to one potentially valued set of activities.

That the RPM measures, and reveals something about, basic cognitive functioning actually follows from the application of Item Response Theory in its construction.

Item Response Theory (IRT) was developed in Britain in the early 1930s, used in the development of the RPM, translated into mathematical formulae by Rasch in the early 1940s (in the course of which he specifically tested his formulations by showing that they applied to the RPM [Rasch, 1980]), and popularised in the US and elsewhere by Wright and others (e.g. Wright and Panchapakesan, 1969) in the 1960s.

To establish the internal consistency of the RPM, graphs (Item Characteristic Curves, ICCs) were plotted (Raven, J.C., 1939) to show the way in which the probability of solving any one item related both to the probability of solving every other item and total score. To the extent that these graphs reveal that the probability of solving any one item does indeed increase in step with the probability of solving easier and more difficult items, it shows that, whatever the manifest content of the items, they are tapping some common underlying continuum.

Given that the manifest content of the items changes from simple perceptual ("Gestalt") items, through easy analogies, to complex analogies which seem to require considerable "analysis" to discern and isolate the "relevant" elements, this shows that "perception" and "reasoning" form part of the same psychological continuum. Put the other way round, perception is not an immediate, visual, process but involves *conceptual* activity. Such activity is required to discriminate figure from ground and relevant from irrelevant. It is therefore a mistake to regard the RPM as

a measure of "problem-solving ability" since, as Spearman (1927) was at pains to emphasise in his principles of noegenesis*, the capacity to identify and handle problems depends on *simultaneously* developing an understanding of the whole in order to know what to look for in the parts (i.e. in order to "analyse") *and* knowing which parts to discriminate from background "noise" in order to "see" the whole.

It follows from what has been said that the item analysis – the set of Item Characteristic Curves – for the RPM demonstrates (a) that something which might tentatively be named "general conceptual ability" does "exist", (b) that the RPM in some sense measures this ability, and (c) that the qualitatively different items of which the test is composed form part of a common continuum. These qualitatively different types of item are not measuring "different things". Just as the concept of "hardness" is not negated by the fact that it is different substances which display different degrees of the characteristic, so the fact that the items in the RPM differ in their manifest content does not invalidate the notion that their solution demands different levels of "cognitive ability". The ability to solve one type of item increases incrementally and in step with the ability to solve other types. There are no metamorphoses in thinking between the ability to solve one kind of item and the next (although this does not imply that there are not spurts and plateaux in individual development).

At this point it is necessary to counter the objection that factor analysts have isolated separate factors made up of these "perceptual", "reasoning", and "analytic" items. I have shown elsewhere (Raven, Ritchie & Baxter, 1971) that the correlation matrix obtained by intercorrelating the items of a perfect Guttman or Rasch scale can be fitted by neither a principal components analysis nor by any orthogonal or oblique rotation. The nature of the correlation matrix is determined by the properties of such scales. A knowledge of whether someone gets a very easy item right does not enable one to predict whether they will get a difficult item right. The correlation between very easy and very difficult items therefore tends to zero. On the other hand, items of similar difficulty are highly correlated: A knowledge of whether someone gets one item right or wrong is a good predictor of whether he or she will get the next most difficult one right or wrong. The correlation matrix obtained by intercorrelating the items after they have been arranged in order of difficulty thus has correlations tending toward unity around the diagonal and approaching zero in the

* The word noegenesis derives from the Greek word *noetic*, and thus means "mind creation".

distal corners. This correlation matrix cannot be re-created by multiplying and adding loadings on *any* set of factors smaller in number than the original items. If one forces data of this kind into a factor analysis one gets a series of "power" factors. These are made up of items of "similar" difficulty because adjacent items intercorrelate highly. (The average within-factor correlation is determined by the number of factors extracted.) But now comes the misinterpretation. Items of similar difficulty consist predominantly, though not exclusively, of items of the same manifest type. In fact, the factors contain some – in reality easier – items from the qualitatively different type which comes developmentally earlier, and some – in reality more difficult – items from that which comes developmentally later than the bulk of the items in the cluster. But these "non-conforming" items can easily be overlooked when naming the factor. Researchers have tended to name these factors to reflect their dominant manifest content when they are, in reality, power factors.

We can now return to our conclusion that the IRT-based item analysis of the RPM really does show that there is a *continuum* in "cognitive (actually 'conceptual') ability" and that this continuum can be assessed using a range of items running from easy "perceptual" items to difficult "analytic" ones. It involves the ability to discriminate figure from ground; the ability to discern order and meaning in (or make meaning out of) confusion; the ability to form high-level, usually non-verbal, concepts which enable one to make sense of the environment. Spearman used the Latin word *educere* – to draw out – to characterise and discuss this component of General Intelligence – ***g*** – and contrasted it with *reproductive* ability, the ability to reproduce already verbalised knowledge.

The conclusion that something which might be termed "general conceptual ability" or "eductive" ability "exists" has been reinforced, and its generalisabilty underlined, in a number of studies in which the RPM has been correlated with other tests.

However, both its existence and generalisability have been neatly confirmed in the study reported by Styles in an earlier chapter in this volume and in Styles (1999). This study, like the scaling procedures used in the development of the tests, was grounded in IRT.

Using a mathematical formulation of IRT, Styles mapped the levels of thought revealed by three Piagetian tasks – the Balance, Chemical Combinations, and Correlational tasks – onto the set of RPM ICCs.

What emerged was that the (Piagetian) level of answers given to these problems increases gradually and incrementally in step with the ability

to solve RPM problems of similar difficulty. It is again clear, therefore, that the ability to solve qualitatively different types of problem develops progressively and does not emerge from the kind of metamorphosis which has sometimes been said to lie behind development of the ability to solve the more complex Piagetian problems.

Styles and Andrich's study not only provides a further demonstration that the RPM is tapping a continuum of fundamental psychological importance, it also indicates that whatever is being measured cannot be dismissed as an ability of mere academic interest. It follows from their work that RPM scores reflect the ability to solve complex, "real-life", problems of an apparently very different character.

A quite different line of work showing that the RPM taps basic psychological abilities comes from researchers working with Reaction-time and Inspection-time (Jensen, Larson & Paul, 1988; Vernon, 1989, 1991, 1993; Deary, 1993, 1995).

Vernon and Deary independently concluded from their reviews of the work of a number of researchers that the RPM has significant, but not strong, correlations with:

- *Low* cerebral glucose metabolic rate. That is, those who get high scores on the RPM appear to work more efficiently.
- *Some* – but far from all – measures of "reaction-time". (Many measures of "reaction-time" do *not* correlate with the RPM, and the studies which have reported the highest correlations have included speed of response to difficult "IQ-type" questions among the measures composited. No *general* statement to the effect that the RPM and "reaction-time" measure the "same thing" is, therefore, justified.)
- *High* cortical response (averaged evoked potential) as measured by EEG, to unanticipated visual stimuli – but low cortical response to self-administered stimuli.
- Blood calcium level, which is itself associated with neural conductivity.

Most impressive, however, is Deary's demonstration that the RPM – and eductive ability more generally – is more highly correlated – at the level of about .6 – with the amount of time people require to be 85% accurate in perceiving which of two lines of very different length is the longer, i.e. "Inspection-time". (Deary makes a point of emphasising that the measurement of Inspection-time does NOT require the person being assessed to work at speed. Inspection-time is measured by varying the

amount of time the lines are exposed and finding the point at which respondents are unable to discriminate accurately between them.)

Inspection-time, like eductive ability, and unlike reproductive ability, "declines" with age ... i.e. has gone up with date of birth.

There is one other recent study which deserves to be singled out for mention here. Carpenter, Just, and Shell (1990) reported that 95% of a sub-set of verbally encoded *Advanced Progressive Matrices* items could be solved by a computer programme which was required to check for the presence or absence of only five rules which might govern the orderliness of the matrix and which, if present, collectively determine the characteristics of the element required to complete the pattern.

Progressive Matrices items were not constructed with a view to analysing the "problem-solving" strategies employed by respondents. As a result, the items of the classical series often have features which make it difficult to identify the operation of the rules Carpenter et al. sought to study. Likewise, the set of options from among which the correct answer has to be selected were not constructed in such a way as to make it possible to test hypotheses derivable from their theory about how the kinds of error which are made should relate to total score (although J.C.Raven did in fact find that type of error was directly related to total score).

Vodegel-Matzen (1994a&b) constructed a set of items which (a) contained no features extraneous to Carpenter and Just's framework that might influence their difficulty, (b) had theoretical difficulty levels which could be calculated from that theory, and (c) had distracters which differed systematically from the correct answer only in terms of the number and kind of rules omitted (and the probability of selection of each of which could therefore be expected, on theoretical grounds, to vary systematically with total score).

This new test had both excellent internal psychometric properties and a very high correlation with the RPM.

The results of the error analysis were as predicted. The most able of those who were unable to solve a given item selected answers which failed to take account of a single – the most difficult – rule governing the orderliness of the matrix. Less able respondents overlooked more rules. Thus the *type* of error made varied with total score in the way predicted by the theory. The finding thus gives us new insights into the causes of deficiencies in cognitive functioning.

Another factor determining item difficulty is the ease with which it is possible to identify the elements to which attention needs to be paid when trying to identify systematic variation between the cells of the matrix.

By making the elements of which the matrices were composed more "life-like" – i.e. using such things as hats, bananas, and faces instead of squares and triangles – while requiring respondents to apply the *same* rules in order to solve the problems, Vodegel-Matzen demonstrated just how important this factor really is. But what was most interesting was that the change to more life-like components made the items easier *for everyone* – not just for lower-scoring respondents. The rank-order of items and respondents remained virtually unchanged.

Use of pictorial elements may, however, result in cultural differences which are not found on the original test.

Going the other way – "hiding" the features that it is important to attend to – making "correspondence-finding" more difficult – makes the items harder for everyone. But it makes little difference to the order of difficulty of the items or the rank-order of respondents.

Moving away from specific studies to programmes of research, remarkable confirmation of the appropriateness of the eductive-reproductive framework for thinking about these abilities has come from an entirely unexpected quarter. Cattell (1963) and Horn and Cattell (1966) initiated a stream of research in the area by proposing that the basic distinction was between "fluid" and "crystallized" intelligence and further suggesting that the latter "differentiated out of" the former. Whereas Spearman argued that the natures of the two abilities were "trenchantly contrasting", Cattell and Horn viewed them as closely related and expected to find that they had a common neurological substrate which, they hoped, would be illuminated by research using "more fundamental" psychophysical and psychophysiological measures.

On reviewing the available material for Sternberg's (1993) encyclopaedia, however, Horn came to a series of conclusions which support Spearman's standpoint in virtually every detail: (1) the thousands of "more fundamental" measures which have been developed do *not* cluster into the eductive vs reproductive domains but generate some eight *additional* factors or components of "intelligence"; (2) none of these additional factors has anything like the explanatory power of eductive and reproductive ability; (3) "crystallized" intelligence does *not* "differentiate out" of "fluid" intelligence; the two are distinct from the beginning; (4) the two abilities have different genetic origins; (5) the two are affected by different aspects of the environment; (6) the two follow different developmental trajectories over the life cycle; and (7) the two predict very different types of performance.

Given this remarkable convergence between what were very different positions, it remains only to ask which terminology seems most appropriate and to suggest that the eductive-reproductive formulation seems less likely to mislead.

Construct Validity: A Measure of Problem-Solving Ability or a Measure of Intelligence?

The RPM was constructed neither as a measure of "problem solving ability" nor as a measure of "intelligence". Nevertheless, many researchers have treated it as if it were a measure of one or the other. In the next few paragraphs the conceptual difficulties involved in doing this will be discussed.

Problem-solving is a difficult and demanding activity. It requires people to be sensitive to fleeting feelings on the fringe of consciousness which indicate that something could be done better or merits exploration. It involves initiating, usually on the basis of "hunches" or feelings, "experimental interactions with the environment" to clarify the nature of a problem and potential solutions. Having used their feelings to initiate activity, people need to monitor the effectiveness of their actions in order find out what is working and what is not, and why. In this way they can learn more (not necessarily consciously) about the nature of the problem and the effectiveness of their strategies. They can then modify their behaviour and launch a further round of "experimental interactions with the environment".

Beyond what may be regarded as *process* components of problem-solving lie a set of social and personal beliefs – beliefs about society, how it works, and one's own place in it. These include the belief that one has a *right to* ask questions and to do such things as try to influence the way society works.

And, in addition to these internal components, effective problem-solving often also involves persuading other people to help, prising information out of other people's heads, and learning how to do things by imitating others.

It is important to note that all this implies that what is often thought of as "cognitive activity" is primarily affective, conative, and interpersonal. Without the use of feelings there would be no insights; without persistence (conation) there would be no testing of those insights; and without

actual behaviour (experimental interactions with the environment or "conversations with the problem") followed by feeling-based monitoring of the effects of that action, there would be a major failure in observation, "thinking", and learning.

It follows that it is not legitimate, except for purely conceptual purposes, to try to separate the cognitive elements of eductive activity from its other components. The process which is commonly described as "cognitive" is necessarily *primarily* dependent on affective, conative, and social processes. The attempt to develop "pure" measures of cognition is doomed to fail because the very basis of the attempt seeks to eliminate the processes on which effective cognition is most dependent. We will come back to the question of how effectively the RPM copes with these problems when we have completed our theoretical discussion.

No one is going to undertake any of the activities discussed in the last paragraph unless they, in some sense, care about the activity. It is difficult to formulate this statement more precisely because the kinds of things people are strongly motivated to do often seem to have much in common with compulsions. People do the things they are strongly motivated to do persistently and repeatedly despite punishment, despite their better judgement, indeed, "despite themselves". Of course, that is a circular statement.

The goals or contexts in relation to which people will undertake difficult and demanding activities like "problem-solving" vary enormously. One person will, for example, engage in them mainly while trying to advance scientific understanding. Another while trying to put others at ease. Another in the course of trying to advance him or herself in a career. And yet another while seeking to control others.

An important implication of this observation is that people will only develop important components of cognitive competence while they are working at tasks they are intrinsically strongly motivated to undertake. This point will not be developed here. A discussion will be found in Raven (1987).

The implication for assessment is that people's ability to carry out the kinds of activity that are needed to identify and solve problems is only likely to reveal itself when they are undertaking activities that are important to them.

This conclusion resonates with the views of "situated cognitionists" like Greeno (1989) and Brown, Collins and Dugoid (1989). However, our conclusion differs from theirs in that we are arguing that *the same*

psychological processes **do** occur in different contexts. They simply *look* different, just as copper looks and behaves differently when situated in the context of sulphur *and* oxygen as compared with, say, just oxygen.

What these observations mean is that the effective assessment of high-level competencies – including problem-solving ability – is dependent on the adoption of a *two-stage* measurement procedure. One must first find out what kinds of activity people find motivating (i.e. "engaging", "important", or in some other sense "valued"), and then, and only then, assess *how many* of the activities which make for effective "problem solving" they display while carrying out those activities.

People have too often been branded as "unable to think" simply because they do not "think" in a mathematics class or while undertaking tasks set by others in an Assessment Centre (or during a Piagetian experiment). Such people may be able think *very effectively* (i.e. make their own observations, learn without instruction, and make good judgements) on a football field, or when developing better materials for building the bridges which are to be assembled by the team in the course of leading which they are supposed to demonstrate their capacity to think in an Assessment Centre. The question which should be asked is, therefore, frequently *not* "How intelligent is this person?" but "*While undertaking which kinds of activity* does this person display his or her intelligence?" Only after that can one meaningfully ask: "Which of the competencies required for intelligent behaviour does he or she display in the course of these activities?"

This psychometric oversight has not only led to injurious and misleading assessments of individuals, it has also resulted in unjustifiable conclusions being drawn from research. These research conclusions have then often contributed to the introduction or perpetuation of damaging educational, occupational, and social practices. Insufficiently *comprehensive* assessments must be regarded as unethical: They have detrimental consequences for the individuals assessed and for others who would have benefited from the educational and other programmes which have been condemned. These detrimental consequences cumulate for society: Individuals who have been mis-assessed are often deprived of opportunities to contribute as they might to society and the cancellation of educational and social programmes which, in reality, have genuinely beneficial effects can have serious social consequences.

A series of seriously misleading "findings" arising from failure to employ appropriate measurement practices will be found in Raven

(1991). One of direct relevance to the deployment of the RPM is that cognitive development "plateaus" in adolescence. This conclusion stems from not having measured "intelligence" while those concerned were carrying out activities they cared about and in connection with which they had had opportunities to continue to develop their powers of reasoning. When the "ability to think" is assessed more appropriately, the available evidence suggests that it increases throughout life (Jaques, 1976, 1989; Kohn & Schooler, 1978).

Turning to the widely held view that the RPM measures "intelligence", one of the most fundamental difficulties is that qualities like "intelligence" and "enterprise" are, as Gardner (1987) and Deming (1993) have also argued, qualities which need to be studied and documented at *cultural*, rather than individual, levels. To advance understanding (i.e. to engage in intelligent activity) effectively, one really needs to proceed on a group basis. One needs a wide range of people who do very different things. Thus one requires some people who are good at each of the following: generating ideas; digging relevant information out of a diverse literature; getting people to work together effectively; discerning patterns in accumulating data; deciding what information to collected to test those insights; using their feelings to notice activities that are likely to succeed; capitalising upon whatever is discovered in the course of an "adventure" initiated on the basis of feelings or "hunches"; putting emerging understandings into words; communicating findings to others; and engaging in the political activities necessary to attract the funds needed to continue the work.

Empirical support for the central claim of the last paragraph comes from the work of Taylor, Smith, and Ghiselin (1963), who showed that effective advance of scientific understanding depends on having *teams* made up of people who are motivated to do very different things and who contribute in very different ways to the overall activity. In a similar vein, McClelland (1961) showed that enterprise and innovation stems from many people trying to do whatever they are doing in new ways. More generally, he found that what happens in a culture is primarily dependent on the shared values of the culture. Most important was whether its members would bring to bear multiple components of competence in order to undertake the kinds of activity they cared about effectively. Kanter (1985) has likewise shown that the innovativeness and survival of organisations depends on *everyone* contributing (through "parallel organisation" activity) in very different ways to a climate of innovation and improvement and on steps being taken to recognise and develop

their diverse talents. Based on his own observations, Deming (1993) has made a similar point. Roberts (1968) and Rogers (1962/83) have observed how innovativeness is dependent on "teamwork" and networks of contacts. Dalziel and Schoonover's (1988) research led them to a similar conclusion. Indeed, even Jaques (1989), while emphasising the need for a steep organisational hierarchy based on "cognitive ability", stresses the crucial importance of managers-once-removed devoting a considerable amount of time to thinking about the talents of subordinates-once-removed and how to place, develop, and utilise them.

"Eductive ability" contributes to the effective performance of each and every one of the activities mentioned.

These points can be reinforced, and additional implications highlighted, by reflecting on the way in which the word "intelligence" is used in phrases like "the (military) intelligence service". This reveals that, despite psychologists' inclination to adopt a reductionist definition of intelligence (as in the assertion that "intelligence is what intelligence tests measure"), what has actually happened is that psychologists have omitted from their measures a great deal that should have been there. They have done society a dis-service by leading parents, teachers, and managers to think that intelligence *tests* capture what they, as laymen, understand by the word "intelligence". They have led the members of these groups to overlook a great deal of what they should have been paying attention to as they sought to think about, nurture, and capitalise upon the talents of their children, pupils, or subordinates.

Generating new insights and understandings ("intelligence") through a military or industrial intelligence service clearly involves making sense of confusing and incomplete information. Intelligence officers frequently cannot know beforehand what to observe and report. They depend on their *feelings* ("intuition") and on recognising an emerging pattern to tell them what is significant. The qualities required to make sense of the incoming information include the ability to seek out, collate, re-interpret, and piece together scraps of unreliable and incomplete information in order to perceive something that has not been seen before and to use what is then perceived to tell them what to attend to and observe next and what to report. The qualities required to do well also include the ability to discern what further information would be required to test initial impressions and the determination to collect that information, perhaps through overt as well as mental "experiment".

What has been said so far amounts to nothing more than a statement that numerous components of what we have called "eductive" ability are

required to work intelligently. However, it also illustrates some of the things that have been missing from most previous attempts to assess eductive ability.

But much more is involved in intelligent activity. The qualities required to establish military intelligence also include the ability to prise information out of other people, the motivation and the ability to do such things as set up and manage networks of contacts to obtain information, the ability to make good judgements about who possesses the sensitivities and persistence to do well in the field, and the ability to supply those contacts with appropriate guidance concerning the kind of information to be sought.

The ability to carry out these activities clearly involves eductive ability. But it also involves many other motivational dispositions and abilities and the effective use of accumulated, specialist, knowledge of military operations, people, and systems.

It follows that, for a group to act intelligently, it is necessary to have a wide range of people who contribute in very different ways to establishing and running a network and who find ways of advancing a wide variety of activities. It is not possible for any one person to be motivated to carry out, and be good at performing, all the activities that are important.

Grid 20.1 has been prepared to make this way of thinking more concrete and to link what is being said here to the more general framework for reflecting about competence that will be summarised below and which has been published in full in *Competence in Modern Society* (Raven, 1984).

What has been said indicates that intelligent behaviour occurs when one has a range of people who are strongly motivated to carry out as many as possible of the activities listed down the left-hand side of the Grid and are capable of carrying them out in a co-ordinated way.

In the course of carrying out their chosen activities, each person needs to exercise as many as possible of the competencies listed across the top of the Grid (and others like them).

The first thing to be emphasised is that what is portrayed in this Grid is *not* a *culture* of intelligence. It is *intelligence* itself.

It follows that, while Jaques is right to emphasise the rarity of the motivation and the ability to carry out (and the societal and organisational importance of carrying out) organisational and societal management tasks involving such things as understanding and influencing opaque, international, socio-economic and socio-physical processes, his failure

to recognise at least some crucial components of the organisational arrangements required for intelligence and innovation has led him to some seriously misleading conclusions.

It also follows that, while Gardner is right to stress the importance of multiple talents, he may need to reformulate his theory of multiple intelligences. Our own theory suggests that the conclusion to be drawn from his observations is that there are many important activities which people may be strongly motivated to carry out and in relation to which they may develop and display high-level competencies. But, while there are also many more of these high-level competencies than psychologists have been inclined to acknowledge in the past, their number may still be relatively limited.

The framework developed here in some ways reinforces, but in other ways draws attention to limitations of, the observations of authors like Richardson (1991), Ogbu (1992), Tharp et al. (1984), and Gallimore (1985). These researchers argue that cognitive abilities will be revealed only when people are undertaking tasks which are meaningful and important to them and that their apparent ability to carry out these tasks depends on their prior opportunity to exercise, and thus develop, these abilities. Unfortunately, these authors mainly dwell on the *dominant* values of the cultural groups they studied and the kinds of "intelligence" evoked or called for in those contexts. They fail to note the variance in valued activities within all cultures. As a result they overlook two important things:

(1) That, if one is to nurture cognitive and other high-level competencies (in the way that effective parents nurture such qualities in their children and managers nurture them in their subordinates [Raven, 1980, 1984; Spencer & Spencer, 1993; Kanter, 1985]), it will be necessary to create individualised developmental programmes which engage with people's motives and thus enable them to practise (and thereby develop) these components of competence.

(2) That, if one is to measure eductive ability more effectively than in relation to a task which *most* people find inherently engaging (as in the RPM), one must first find out what kind of activity the person being assessed is strongly predisposed to undertake and then which of the components of competence that are needed to carry it out effectively are displayed while it is actually being carried out. (Instead of doing this, most critics of conventional measurement – such as Piaget, Vygotsky,

GRID 20.1

A MODEL OF INTELLIGENCE

Examples of Competencies Required to Carry out Activities Crucial to Intelligence (Observable only while activities which are personally engaging are being undertaken)

Examples of activities required to create a Culture of Intelligence or Enterprise but which people may or may not be strongly motivated to carry out.	Eductive ability (itself having cognitive, affective, and conative components and involving such things as the ability to initiate and learn from "experimental interactions with the environment").	Reproductive ability: The store of information and intellectual skills available from the past.	Ability to persist.	Ability to use feelings to initiate action, monitor the effects of the action, change one's behavior accordingly, and start a further cycle.	Ability to persuade others to help.	Ability to resolve value conflicts and to integrate values with each other and work toward their achievement over a long period of time.
Tendency to understand and influence the workings of society around the organization - including what is happening on the other side of the world.						
Tendency to generate new formal theories e.g. in connection with the workings of the organization or in connection with technology.						
Tendency to engage in organizational development activity.						
Tendency to notice new things that need to be done.						
Tendency to translate new theoretical understandings into a product.						
Tendency to get people to work together effectively.						
Tendency to think about, place, develop, and utilize the talents of subordinates.						
Concern to put others at ease.						
Tendency to soothe interpersonal tensions.						
Tendency to get together with others and set up indirect strategies to influence people higher up in the organization.						
Tendency to provide help and encouragement to those engaged in innovation.						
Tendency to initiate the collection of, seek out, sift, and come to good innovative decisions on the basis of forward-looking information						

and Richardson* – have simply confronted respondents with a set of problems geared to an alternative, but still single, value system.) The only researchers who have seriously addressed this problem are those who have worked in what may be termed the McClelland tradition.

We may now return to the question of whether, and how, the members of hierarchically differentiated occupational groups within organisations need to differ in motivation and eductive ability for the organisation to function most effectively. Reflection on Grid 20.1 suggests that it may be more important for people working at different levels in an occupational hierarchy to differ in the kinds of activity they are strongly motivated to carry out than in their eductive ability. As Hogan (1990) and Hope (1984) have shown, managers who apply their eductive ability mainly to advancing themselves in their careers (by, for example, getting rid of all personnel who are concerned with future development so as to present themselves as being able to run organisations which are "lean, mean, and profitable" in the short-term) can have disastrous effects on their organisations.

It is equally obvious from Grid 20.1 that, to carry out important valued activities effectively, many other components of competence besides eductive ability are required. The components of competence listed across the top of the Grid (and others like them) are unlikely to be highly correlated with each other. Instead, they contribute cumulatively and substitutively to effective performance, rather like the terms of a multiple regression equation. Competence is a value-based, internally heterogeneous, quality. Its measurement therefore cannot be assimilated to the internal-consistency model which dominates mainstream psychometrics.

Despite the implications of what has been said, it is obvious that we need to develop a more adequate descriptive framework to help us think about the components of competence listed across the top of the Grid. At present, for example, the use of feelings and persistence appear both as

* Unlike those of Vodegel-Matzen, Richardson's "real-life" matrices do not exhibit the same logical operations as the diagrammatic matrices with which they are said to be isomorphic. They cannot be of equivalent difficulty because they do not have as many transformations going on at the same time (cf. Jacobs & Vandeventer, 1968) and do not exhibit serial change of the same order of complexity in two dimensions simultaneously. They do not have the properties of mathematical determinants because the argument that applies in one direction does not apply in the other. Perhaps still more importantly, they do not require respondents to *simultaneously* attend to their emerging understanding of an overall pattern in order to discover what to pay attention to in the parts and to attend to the parts in order to discern the overall pattern, i.e. they do not require the same degree of *meaning-making* ability.

components of eductive ability and as qualities which make an important independent contribution to effective behaviour. We also need a better framework for thinking about the potentially valued styles of behaviour that appear down the side.

An examination of Grid 20.1 helps us to understand how the abilities assessed by the RPM contribute both to a wide range of occupationally relevant performance *and* to some occupationally dysfunctional behaviours. However, it also helps us to understand why the RPM:

- Does not necessarily reflect the level of eductive ability which people are capable of displaying *while carrying out tasks they care about*.
- Does not correlate more highly with occupational performance: Occupational performance is determined by whether an individual's values are aligned with those required to perform the job effectively, by the possession or otherwise of numerous other competencies, and by what other people do.
- Could probably, through a series of precisely targeted studies, be shown to be much more highly correlated than currently appears to be the case with the ability to carry out each of a wide range of important activities.
- Does not correlate more highly with level of job attained and retained. As things are currently organised in Western cultures, one would expect this to be more strongly determined by a valuation for personal advancement than by competence at doing the things which those employed in any position need to do to improve the overall effectiveness of the organisation.

It also helps us to see that intelligent behaviour involves an extended time dimension that is commonly overlooked, especially during assessment. To behave intelligently, one must organise one's life in a such a way as to be able to achieve one's valued goals effectively. To do this it is necessary to bring to bear relevant past experiences, imagine potential future scenarios, anticipate obstacles to their achievement, and find ways round the obstacles. It is necessary to resolve value conflicts, among other things by considering the probable consequences of alternative courses of action. The consequences to be considered run from personal (individual) consequences, through organisational consequences, to societal consequences. To consider the last two it is necessary to build up one's own understanding of social and ecological processes. To enact the conclusions of such reflections it is necessary to take a stand for what

one believes to be not only in one's own long-term best interests, but also those of one's family, organisation, community, society, and planet. It is these connections which result in cognitive ability being psychologically bonded to a valuation for such things as taking responsibility for others and taking one's own moral decisions. And they also explain why the adoption of reason-based discipline strategies results in the enhancement of eductive ability.

From a practical point of view, it is clear from Grid 20.1 that using the RPM as a selection and placement tool without the simultaneous use of more broadly based measures is inadequate because many people do not apply their eductive ability to doing what others need them to do. This observation underlines the importance of pressing, not only for developments in *assessment,* but also for more studies of what the short and long-term, personal, organisational, and societal consequences of people doing different things actually are. Given such information, we would be able to generate more meaningful job specifications.

It is also evident from Grid 20.1 that undue reliance on selection procedures which claim to identify "highly able" people may have the effect of absolving teachers and managers from two of their primary responsibilities. These are, on the one hand, to create developmental environments and, on the other, to introduce guidance, placement, and development activities which will help to develop, utilise, and recognise the contribution of, people who value, and are able to undertake, all of the activities revealed by considering each of the cells of the Grid.

To help society and organisations tackle this problem it is important for psychologists to engage in a number of different activities. They must help teachers, managers, and society to clarify the activities which may need to be carried out, develop the tools which are required if assessment systems which recognise and capitalise upon wider aspects of competence are to be introduced, clarify the organisational arrangements which are required if the results of staff and organisational appraisal activities are to be fed to audiences who will help to ensure that action is taken, and develop the understandings required if teachers and managers are to create developmental environments and climates of innovation which will enable society to develop and utilise all the human resources that are available. Developments in all these areas are vital if we are to reduce the most widespread and most serious misuses of tests highlighted by Raven (1991) and Moreland et al. (1995). Preliminary work to help fill some of them is summarised in Raven (1984, 1994).

Predictive Validity

Although the RPM was developed for research purposes, it is widely used in psychological practice for selection, guidance, and problem diagnosis and remediation. An examination of its predictive validity is therefore called for. As it happens, this will throw further light on its construct validity.

Educational Success

Numerous studies (see Court & Raven, 2001; Court & Raven, 1995) have shown that the RPM correlates with school performance, although, as the theoretical basis of both tests would lead one to expect, the correlations between school performance and the *Mill Hill Vocabulary* (MHV) test – a measure of *reproductive* ability – are generally higher than those with the RPM.

Unfortunately, these correlations do not exactly provide a cause for jubilation because, as will be suggested in the next paragraph and as the author has argued more fully elsewhere (Raven, 1991), most measures of educational performance themselves lack construct validity.

Consider the typical "science" test. There is no sense in which such a test assesses competence to function as a scientist, whether in a scientific career, in other fields of work, in the home, in politics, or in the community. The competencies required by scientists include the ability to problematise, the ability to invent ways of collecting relevant information, the ability to locate appropriate mathematics or other ways of summarising data, the ability to persuade others to collaborate, the ability to work with others, and the ability to communicate. In no sense does the typical science test assess such things. Instead, it measures the ability to present temporary knowledge of minuscule and arbitrary selections of out-of-date information (which also has little chance of relating to the assessee's current or future needs) in a way that meets the examiner's expectations (which themselves typically embody an inappropriate concept of science). Such tests, measuring neither scientific competence nor a knowledge of "science", clearly lack construct validity. (Given this understanding of what they actually measure, it is not surprising that they correlate more highly with measures of reproductive ability than with measures of the ability to make meaning out of confusion; the ability to perceive and think clearly.)

Similarly, tests of "English" (and, by implication the "ability to communicate") which ask students to do such things as underline the

verbs in sentences also lack construct validity. Effective communication involves the deliberate manipulation of structure to create and convey an impression, the use of allusion to evoke emotions, the use of innuendo and the evocation of feelings to elicit behaviour, and the ability to write with sensitivity to the values and prejudices of a target audience in order to induce desired action.

These examples highlight two of the problems that are inherent in conventional ways of thinking about the procedures to be adopted when establishing test validity. They illustrate the *criterion* problem in the very field – educational testing – in which testing is most widely applied. Yet, as McClelland (1973), Messick (1989), and others have shown, the problems become more numerous and more serious as one moves into the field of occupational testing.

Occupational Success

The technical and logistic problems involved in establishing the predictive validity of a test in occupational settings include:

a) *Problems associated with the Criteria of Success* (including their validity as indices of the construct being assessed): The qualities apparently required to perform a job "well" depend on the criteria adopted when evaluating performance. Different qualities are, for example, required to secure rapid advancement in an organisation, to secure the survival of that organisation through the invention of new products, to secure its growth through financial and/or political manipulation, and to secure the survival of society. Those who are best able to obtain the esteem of those above them are not necessarily best at releasing the energy and talents of their subordinates and, indeed, often advance themselves by applying their cognitive abilities to make their sections appear more "efficient" by getting rid of the personnel, the time, and the networks of contacts which are required for institutional development, and by eliminating those with alternative viewpoints who might challenge their views or compete for their position (Chomsky, 1987; Hogan, 1990, 1991; Hogan et al., 1990; Jaques, 1989; Nuttgens, 1988; Raven, 1984; Raven & Dolphin, 1978; Spencer & Spencer, 1993).

b) *Problems deriving from the use of Inadequate Job Analyses and Job Descriptions*: The activities required for the effective performance of a job may differ from those identified in the job

description and thus be overlooked when attempts are being made to validate selection procedures (Fivars & Gosnell, 1966; Klemp, Munger & Spencer, 1977; McClelland & Dailey, 1973, 1974; Raven, 1984; Spencer & Spencer, 1993; Taylor & Barron, 1963). Indeed, the notion of "effective job performance" is itself problematical. Thus bus driving can be construed as involving only such things as the ability to avoid accidents. Yet, as Van Beinum (1965) has demonstrated, the effectiveness and adaptation of a bus service is dependent on bus drivers sharing their insights with their managers and contributing to a climate of innovation. Kanter (1985) has generalised the point: the innovativeness and survival of organisations depends on people doing things which would never be suspected if one asked merely "What do they need to do to produce widgets?"

c) *Problems created by the use of Inappropriate Selection Procedures in the past*: Those best able to perform a job may have been (intentionally or unintentionally) eliminated from those admitted to the workforce. If this has happened it will be impossible to demonstrate the importance of the required qualities (Berg, 1973; Holland & Richards, 1965; Hope, 1984; McClelland, 1973; Raven, 1994; Raven, J. & Stephenson, J. (Eds.), 2001; Taylor, Smith & Ghiselin, 1963).

d) *Problems created by the Non-Attributable Nature of Outcomes*: In most organisations it is extremely difficult to attribute observable effects to any one person or group of persons (see Day & Klein, 1987). This is especially so when circumstances are continuously changing and the effects of actions may take many years to show up. This makes it difficult to collect accurate information about whose work genuinely benefits an organisation and distinguish those who confer important benefits from those who are only able to create a good impression and move on before their mistakes are discovered.

e) *External Constraints*: Organisational arrangements, and other people's expectations, may prevent people doing the things required for effective job performance.

f) *Change Over Time*: People do different things in the "same" job at different times. They may, for example, engage in routine activities for part of the day and in innovative ones at other times. They may develop technological innovations early in their careers

and engage with the political processes which control the funding for such innovations later in their lives.

Despite these problems, it has been shown that the RPM *does* relate to a variety of measures of managerial performance: Staff and financial turnover, profitability, and the ability of the firm to survive financial and other crises. Thus, data supplied by a transnational corporation which runs several thousand small retail stores shows that the RPM correlated .50 with the *Watson-Glaser Test of Critical Thinking* and .20 with assessments of work management performance, .13 with assessments of interpersonal skills performance, and .12 with planning and problem solving performance. Although these correlations are statistically highly significant (being based on a study of 1120 managers), their true significance only emerges as one realises that the correlations between the performance measures and most of the other tests used in the study were zero.

Ingleton (1990) found that while managers with high vocabulary test scores performed well in unchanging conditions, it tended to be those with high RPM scores who were best able to help their firms weather the crisis produced by the 1970s oil price increases. (It is for reasons like this that it is so important for the public service, in particular, to recruit, and to promote into positions having very different job descriptions, a wide range of people who have distinctive patterns of motivation and ability.)

But it is not only in management settings that the tests have been validated. Several studies (see Court & Raven, 1995) have shown that the RPM and MHV between them can predict about 10% of the variance in performance within a wide range of occupations. Validity generalisation analysis, which adjusts these figures for restriction of range and the unreliability of criteria, suggests that a "truer" estimate of the proportion of variance accounted for is 25%. However, since those concerned with personnel selection are necessarily operating in situations involving restricted range and unreliable criteria, it is not entirely clear that the adjusted figure conveys an appropriate impression of the benefits that can be obtained from testing.

One unpublished study illustrating the use of the APM in predicting non-managerial performance involved computer programmers. The data (supplied in tabular form) showed that the APM, administered without a time limit, was a particularly good predictor of success. This is perhaps because similar levels of attention to detail, checking, and persistence are required for success at both tasks.

More generally, meta-analyses (Ghiselli, 1966; Hunter & Hunter, 1984) show that tests of intellectual ability predict proficiency within at least the following types of work: managerial, clerical, sales, protective professions, service jobs, trades and crafts, vehicle operation, and simple industrial work.

All such studies yield what may be regarded as relatively low predictive validities. There is, however, another way of coming at the question of validity which yields a much more positive conclusion. Instead of seeking evidence for the predictive validity of the RPM *within* occupational groups, one can focus on its ability to predict the *level* of job an individual will attain and retain.

Before discussing this topic further, it is necessary to examine more carefully the nature of the activities which distinguish more from less effective performance both within and between jobs.

Critical-incident studies (such as those summarised in Raven, 1984, Spencer & Spencer, 1993, and Raven & Stephenson, 2001) have shown that effective performance in a wide range of jobs depends on doing such things as building up one's own understanding of the way in which the organisation in which one works functions, viewing one's own part in it in appropriate ways, taking initiative to intervene in organisational processes when necessary, building up one's own understanding of the workings of external political and economic systems and intervening in them for the benefit of one's organisation and society, and thinking about the motives and talents of subordinates and how best to place them so as to harness their motives and develop their talents.

Although, as shown in data summarised in Raven (1984), more effective workers in all occupations are distinguished from their less effective peers by the frequency with which they do such things, Jaques (1976, 1989) has argued that these high-level activities are more important in high-level jobs.

He has also argued that the ability to undertake many of these activities is primarily dependent on "cognitive ability". However, he defines "cognitive ability" to include the use of feelings to initiate action which is then monitored to learn more about the situation with which one is dealing and the effectiveness of the strategy one has adopted – together with the ability to take corrective action when these observations show that it is necessary. Such activities require great determination and persistence. Precisely because Jaques wishes to include these affective, conative, and "experimental action" components in his concept of "cognitive ability" he

denies (as we did above) that "intelligence" tests measure it. Nevertheless what he has in mind does seem to have much in common with "eductive ability" as identified by Spearman and as conceptualised here.

Note the problems Jaques' contentions pose for test validation. Even supposing we had a test which adequately measured what he means by "cognitive ability", we would need a collection of very sophisticated studies to validate it. To get high zero-order correlations between the test and criteria it would be necessary to find an organisation in which people were not constrained by day-to-day pressures to attend to matters that did not require them to exercise the maximum level of cognitive capacity of which they were capable. The organisation would also need to be one which *did* require them to apply their cognitive ability to undertaking the kinds of activity mentioned above and which discouraged them from applying it to such things as securing their personal advancement mainly by creating a good impressions on their superiors without doing the things that needed to be done. Alternatively one would have to make a series of detailed – almost ethnographic – studies of what individuals were actually doing in their jobs and relate test scores to conceptually crucial components of that performance.

These observations strongly reinforce the claims of McClelland (1973) and Messick (1989) that the validity of a test cannot be estimated directly. An impression of its validity can only be achieved by first making a theoretical analysis of what the test measures, the competencies required in particular types of job, and the organisational arrangements through which work is conducted, and thereafter reviewing studies – each imperfect in itself – which illuminate what the test measures and predicts and the factors which enhance or reduce the observed relationships. Thus test validation involves nothing less than applying (properly understood forms of) scientific method to illuminate a hidden reality (House, 1991). It is more than a little unfortunate that Barrett and Depinet (1991) do not seem to have understood this position when preparing their highly influential, but altogether misleading, paper.

With these reservations in mind, we will now review evidence suggesting that the RPM, and other measures of ***g,*** are better at predicting the *level* of job an individual is able to attain and retain than at predicting performance within any particular occupation.

Vernon and Parry (1949) summarised the results of testing 90,000 British naval recruits with a short, non-cyclical, version of the SPM during the Second World War. There were systematic differences in the mean

scores of men from 12 general classes of occupation: clerical, electrical workers, precision workers, woodworkers, sheet metal workers, machine operators, retail tradesmen, building workers, "mates", drivers, farm workers, and labourers.

Foulds and Raven (1948) tested the entire workforce of a photographic factory and found very large average differences in the SPM scores of workers at five different levels (Table 20.1).

de Leeuw and Meester (1984) showed that about 50% of the variance in occupational level can be predicted from RPM scores.

Fraser-Roberts (1943) likewise found that there was a marked correlation between RPM scores and level of job attained and retained.

If Jaques is right to argue that cognitive ability is closely related to level of job attained and retained, there ought to be an *optimal* range of scores – neither too high nor too low – for most jobs. The most convincing evidence on this point comes from the work of Hope (1984) which will shortly be reviewed in some detail. However, evidence supporting the argument that there is an optimum range of scores for each occupation comes from a study conducted by J.C. Raven and his colleagues (Crichton Royal, 1957) among telephone engineers.

The conclusion that the relationship between ability and performance is curvilinear may be reconciled with the finding of Hunter and Hunter (1984) that the relationships within all groups are linear by recalling the criterion problem. In Raven's study, the finding was not that higher scoring employees performed worse but that higher and lower scorers left the employment. This strongly supports Jaques' contentions.

But while Jaques' argument is plausible, Kanter's work and the previously mentioned studies showing that eductive ability is important at all levels – especially when the criteria applied in test validation include the survival of the organisation concerned or the society in which it is located – suggests that it is not the whole story.

Social Mobility

Despite the absence of a dramatic relationship between most psychological tests and measures of work performance, the RPM, and "intelligence" tests in general, do not do a bad job of predicting social mobility. Unfortunately, this is again not quite such a cause for celebration as might at first sight appear. The problem is that the link between "cognitive ability" and social mobility is not necessarily direct and may be via patterns of motivation. Given the limited data currently available, it is impossible to

decide whether the link is indeed direct (as Jaques would have us believe) or whether motivational predispositions are responsible for both the test scores people attain and their social mobility.

Perhaps the most impressive evidence of the power of "intelligence" tests to predict social mobility comes from the *Scottish Longitudinal Mental Development Survey* (Scottish Council for Research in Education, 1933, 1949, 1953; MacPherson, 1958; Maxwell, 1961, 1969; Hope, 1984).

Using these data and others, Hope (1984) showed that (a) some 60% of social mobility (both upward and downward) in both Scotland and the US can be predicted from 11 year olds' intelligence test scores; (b) that, by the time children are 11 years old, Scotland achieves (or did achieve) a degree of association between "intelligence" and socio-economic status (SES) that is not achieved in America until age 40; and (c) that, even when the effects of home background are partialled out, children's "intelligence" makes a major contribution to a variety of indices of their occupational success at 28 years of age. The contribution of intelligence is very much greater than that of educational achievement and, as the slow sorting process in America makes clear, is not a surrogate for sociological tracking by the educational system. Early success in the educational system predicts later educational success – but success in the educational system has very little predictive power outside. On the other hand, "intelligence" and, importantly, teachers' ratings (at age 11) of qualities like originality, creativity, determination, and persistence independently enable one to predict life success.

So far so good. The problem is that children from the same family vary almost as much in the kinds of activity they are strongly motivated to carry out (or can be said to value) as in their "intelligence" (Kohn & Schooler, 1978; Raven 1976, 1977), and the available evidence suggests that social mobility, both upward and downward, can be predicted every bit as well from a knowledge of the activities they are strongly motivated to carry out as from their "intelligence". Kohn (1969/1977; Kohn et al., 1986) among others demonstrated that people occupying high socio-economic status positions in several different societies embrace activities like thinking for oneself, originality, taking responsibility for others, and initiative. In contrast people occupying low socio-economic status positions stress toughness, strength, obedience, and having strict rules and moral codes to guide their lives. Kohn initially believed that these differences were a product of occupational experience (and, indeed, to some extent, they are). But, by sectioning the data we obtained from adolescents by

Table 20.1. Standard Progressive Matrices
Score Distributions for Five Classes of Employee in a Photographic Works

	Quartiles of score distribution			
	1	2	3	4
Directive and Executive	79	9	12	-
Highly skilled workers	48	23	19	10
Skilled workers	29	25	27	19
Qualified workers	18	26	28	28
Unskilled workers	12	15	28	45

From Foulds and Raven (1948)

origins and anticipated occupational destinations, we (Raven et al., 1975; Raven, 1976) were able to show that there was a great deal of variance in the concerns of children from similar backgrounds, and that this variance was related to the status of the jobs they expected to enter. This finding, like the finding that two thirds of the variance in "intelligence" test scores is within-family variance, raises serious questions about its origins. A somewhat similar finding was reported by Kinsey (1948). Kinsey found that there was huge variation in the sexual behaviour and attitudes of children who came from similar backgrounds and that this variation predicted where those children would end up. They *joined* others who thought and behaved similarly. Children could hardly have learned sexual attitudes and behaviours so different from those of their parents by modelling or formal instruction. So, where does the variance come from and how does it come about that personal attitudes and behaviour of the kind exemplified by sexual behaviour come to correspond to those of the socio-economic groups people eventually enter? The variance between children from the same family has often been attributed to genetic factors, and, in this context, we may note that Tellegan et al. (1988), Bouchard and McCue (1990), Bouchard (1991), and Waller et al. (1989) have shown that many values and beliefs – including religious beliefs – are as heritable as "intelligence". But, if these attitudes and behaviours are not learned at work and in society, how does it come about that, in the end, their attitudes and behaviours tend to be characteristic of the groups with whom they end up living and working?

Note the problems which these observations pose for the validation and interpretation of "intelligence" tests: Children from similar backgrounds, including members of the same family, vary enormously in both their motives and values and their "intelligence". The variance in their motives predicts their future position in society every bit as well as does their "intelligence". Which is the more basic set of variables? How does variance in "intelligence" come to be linked to variation in motives, values, and personal behaviour?

One study which throws light on the last question has been reported by Maistriaux (1959). Presenting his results in tabular, rather than correlational, form, Maistriaux documents a remarkable relationship between "intelligence" and the kind of activity that people say they want to carry out and enjoy carrying out. Those with higher RPM scores find "intellectual" activities more enjoyable while those with lower scores are more attracted by "practical" activities. In a sense, these results suggest that we may be dealing with different perspectives on "the same" psychological variable.

Other studies – such as those reported by Flynn (1987) and McClelland (1961) – do not, however, support this contention. These studies show that the differences in actual life performance of different ethnic and religious groups in America are very much greater than, and cannot be explained by, differences in their "intelligence". In other words differences in motives, values, and such things as social support, associated with ethnicity and religion are better predictors of "real life" performance than "intelligence".

The overall conclusion to be drawn from this material is that we do not, at present, know whether the portion of the variance in social position and performance that can be predicted from "intelligence" is the same as that which can be predicted from motivation and values or whether the two are additive. In the current state of our knowledge, one clearly has the option of concluding that we should be focusing on the variance in the kinds of behaviour to which people are attracted and their ability to undertake those behaviours effectively rather than on their "intelligence".

Further evidence that the link between "intelligence" and social status and social mobility may be mediated by the kinds of behaviour which attract people comes from two other sources.

The first of these comes from studies of the links between cognitive activity and values. In the first place, cognitive ability and activity is

not universally valued. Many parents do not want their children to ask questions or to be able to use books to find information for themselves (Raven, 1980). Secondly, cognitive ability is psychologically bonded to other personal characteristics, such as curiosity and independence. These may not be valued even if cognitive activity itself is valued (Maistriaux, 1959; Raven, 1987). Thirdly, nurturing cognitive ability depends on child-rearing, educational, and staff-development practices which may not be valued even if cognitive ability itself is valued. Thus the development of cognitive ability is facilitated by the adoption of democratic discipline strategies, encouragement of adventurousness and independence, and studying children's needs and responding to them (Raven, 1980, 1987, 1989a; Feuerstein et al., 1990; Sigel, 1986). It develops in the workplace if managers encourage their subordinates to participate in establishing, and finding ways of achieving, organisational goals and if they study subordinates' motives and talents in order to find ways of developing and utilising them (Kohn & Schooler, 1978, 1982; Jaques, 1976; Lempert, 1986; Lempert et al., 1990).

The second comes from neuropsychology. Trevarthen (1990, 1992) and Sperry (1983) have suggested that the most important psychological concomitants of neurological differences lie in the affective and motivational area. They suggest that the differences in cognitive performance that are associated with neurological locale (including the left and right brain) are merely expressions of more basic differences in motivational predispositions and that it is these which are neurologically located. If "cognitive ability" were assessed while people were undertaking a task that tapped very different motives, such as putting others at ease, not only would our estimates of the "cognitive ability" of those concerned be very different, those abilities would appear to have very different neurological locations. They suggest that the way to make sense of such results would be to recognise, as we have done here, that important components of competence (including eductive ability) will only be displayed while people are carrying out activities they care about. More consistently interpretable data would be obtained by attending to the neurological localisation of motivational predispositions. In that context, Trevarthen underlines the importance of developing a framework for thinking about what he terms the modules of motivation (cf. Murray and McClelland). He moves on to emphasise the need for engagement between the motivational predispositions of parents and children, teachers and pupils, and managers and subordinates, if the development of high-level, generic

competencies, including eductive ability, is to be facilitated (cf. Feuerstein et al. 1990; Vygotsky, 1981; Raven, 1989a).

We may now attempt to draw some tentative conclusions from this review.

1) It is impossible, on the basis of the evidence currently available, to decide whether to explain the allocation of people with different concerns and levels of "cognitive ability" to different socio-economic groups by reference to variations in patterns of value-based competencies or by reference to "cognitive ability".
2) It is impossible to discover whether the relationship between neuro-anatomy and psychological dispositions is to be attributed to differences in motivational predispositions or "cognitive ability" using tests developed within the dominant psychometric tradition.
3) It is particularly difficult to reconcile two sets of claims.
   On the one hand it is argued that:
   - People employed at different occupational levels differ markedly in both cognitive abilities and values.
   - To have an effective organisation it is necessary to have a steep differential in cognitive ability by occupational level.
   - The ability to understand, and find ways of intervening in, the operation of international socio-politico-economic systems for the long-term good of the organisation and the future of humankind calls for exceedingly rare levels of cognitive ability.

   On the other hand it is argued that:
   - The effective performance of low status jobs demands high-level competencies.
   - The culture of intelligence and innovation needed for the development and survival of an organisation or society requires those involved *at all levels* to exercise high-level competencies.

A Framework for Thinking About Competence

Having illustrated some of the limitations of the mainstream "ability" position, the problems associated with the psychometric and validation paradigm with which it is associated, and the vital need to develop a more comprehensive and psychologically appropriate form of assessment, it

is time now to present a brief outline of an attempt to develop a more fruitful way of thinking about competence and its assessment.

But have not numerous psychologists – such as Guilford (1977), Gardner (1985, 1991), Hatch and Gardner (1986, 1990), Sternberg et al. (1986), and Taylor (1971, 1976) – tried to develop such a framework, and have not people like Spearman (1927), Eysenck (1953), Hunter and Hunter (1984), Matarazzo (1990), Barrett and Depinet (1991), and Ree, Earles, and Teachout (1994) shown that all these abilities reduce to the very eductive and reproductive abilities we have been talking about and that no measures of other abilities are both sufficiently distinct from these abilities and sufficiently reliable in themselves to stand up to scrutiny? Indeed they have. Unfortunately, all of these researchers have approached the problem with what might, for the want of a better phrase, be called something approaching a classical psychometric mindset.

Fortunately, some other psychologists have come at the problem from another starting point. Instead of starting with psychometrics, they have studied the nature of occupational, civic, and parental competence. Following Flanagan (1949, 1954), those who have worked in the occupational area have asked supervisors, subordinates, and job encumbants to describe actual incidents of effective and ineffective behaviour – what happened, what led up to it, what the outcome was, what they were thinking and feeling and doing, what other people did, and how others reacted.

Spencer and Spencer (1993) summarise more than 350 studies of this sort, using them to guide their development of a "dictionary" of occupational competencies.

In seeking a way forward here we may first recall that we have seen that, in reality, we need to employ a two-stage measurement model to assess the wider aspects of both "intelligence" and "competence": We first need to discover what kinds of activity people are spontaneously motivated to undertake and then which components of competence they display when undertaking those activities. This means that it will be necessary to develop an agreed conceptual framework for describing the kinds of activity people may "value" and the components of competence they may display while undertaking those activities. [Attempts to develop such a framework have been published by Raven (1984), Huff et al. (1982), and Spencer & Spencer (1993), although Raven and Stephenson (2001) contains a critique of the latter.]

The framework we have ourselves constructed out of that developed by McClelland et al. (1958) for scoring their *Test of Imagination* may be

represented for heuristic purposes in the form of a two-dimensional grid – Grid 20.2 – which is a modified version of that published in Raven (1984, 1991). This lists a number of activities an individual may be strongly motivated to undertake across th e top and a number of the cognitive, affective, and conative components of competence he or she might utilise to carry out those activities effectively down the side.

To move toward a comprehensive assessment of an individual, one could insert ticks (or crosses) in the cells of an extended version of this Grid to show *which* components of competence he or she displayed whilst undertaking each of the activities he or she cared about. One could then reduce data overload by summing the ticks in each column, and compositing the totals for the columns belonging to the Achievement, Affiliation, and Power clusters. This would a yield 3-score, value-based, internally-heterogeneous, personal profile which would be isomorphic with McClelland's *need* Achievement, *need* Affiliation, and *need* Power "motive" profiles. These scores (which obviously have little in common with internally-consistent factor scores) can be understood as being something like multiple-regression coefficients predicting the success with which someone would be able to carry out activities he or she valued.

Despite the succinctness and value of such profiles, examination of the detailed information contained in a completed Grid is much more informative than a collection of scores . This is partly because people may value – or be somehow motivated to undertake – many activities which do not fall into the Achievement, Affiliation, and Power categories and partly because there are many more components of competence than are taken into account in McClelland's scoring system. From a completed Grid one can see *which* competencies the person being assessed tends to display whilst carrying out *which* valued activities.

If one follows the line of argument advanced here further, however, one finds oneself moving away from a concern with scores and *variables* and instead making *descriptive statements* about the people one is assessing. One starts using descriptors (analogous to those used by chemists) to record the activities people value and the competencies they display while undertaking those activities.

However, as soon as one starts to do this, one is forced to recognise that the competencies people will develop and display are in part determined by the extent to which the environment in which they have in the past lived and worked, and the environment in which they are now observed, engages with their values and has led them to develop, and now

leads them to display, the competencies they possess. As a result, one finds oneself attempting to write statements *about those environments* at the same time as making statements about the individual. One then finds oneself trying to say something about the transformations in competence which a change of environment would be likely to effect. One then finds that one has unexpectedly solved the problem – highlighted by Jackson (1986) – of modelling the *transformational* processes which occur in homes, schools, and workplaces.

Operationalisation of the Measurement Model

Those who wish to go into the way in which this framework for thinking about competence and its assessment can be operationalised should refer to Raven (1984, 1988, 1991) or Raven & Stephenson (eds) (2001). Suffice it to say here that there are two main ways in which this can be done. The first involves creating developmental environments which enable people to undertake activities they care about and, in the process, develop and display high-level competencies. The second involves getting inside people's heads in order to find out what motivates them and which components of competence they bring to bear to achieve their valued goals effectively. The latter can be done using specific types of projective methodology, *Behavioral Event Interviewing*, or value-expectancy-instrumentality methodology. As a brief antidote to Barrett and Depinet's failure to examine such methods with any care, the next three paragraphs summarise what each involves.

a) *Observation*. Just as a chemist needs to be familiar with atomic theory to appreciate the significance of a precipitate in a test tube, so the interpretation of what is revealed by behaviour in particular situations is dependent on familiarity with an appropriate interpretative framework. A pre-requisite to eliciting behaviour which reveals which competencies an individual is able to display is the creation of a "developmental environment" (Raven, 1984, 1989a, 1991; Burgess & Adams, 1980, 1986; Stansbury, 1980) which taps the individual's values and leads him or her to display high-level competencies. Thereafter, thorough familiarity with an extended version of the competency framework developed above is necessary to guide the analysis of that behaviour and understand its significance. (It follows that the current drive for

"portfolio" and "authentic" assessments, well-intentioned though it is, is almost certain to founder because of the absence of an adequate descriptive framework for summarising the material.)

b) *TAT and BEI Methodology*. Those scoring McClelland's *Test of Imagination* and *Behavioral Event Interview* protocols follow a detailed and explicit procedure (McClelland, 1951; McClelland et al., 1958; Winter, 1973). Those scoring *Test of Imagination* protocols first ask themselves "What kind of activities is the person who wrote this story motivated to undertake? (i.e. which kinds of activity does he or she value, care about, or somehow feels internally driven to undertake?)", and then "*How many* of a specific and experimentally-derived list of cognitive, affective, and conative components of competence does this person tend to engage in spontaneously while undertaking these activities?" Actually, the process is somewhat circular since a person's motives or values are identified by examining the kinds of things he or she tends to turn thoughts, feelings, and effort into achieving. Nevertheless, the effect is to produce profiles of value-based, internally-heterogeneous, scores of the kind outlined above. *Behavioral Event Interviews* substitute accounts of real-life events for projective stories. One asks people to think of specific times when things were going well (or badly) for them, what led up to the situation, what they were trying to do, what they were thinking and feeling, what they did do, what others did, their reactions to what others did, and what the outcome was (McClelland, 1978; Spencer & Spencer, 1993). More pointedly, they can be asked the same questions about critical life-events identified using Flanagan's "critical-incident" methodology. In the course of these interviews, people's preoccupations (or values) and the competencies they bring to bear to undertake these activities effectively become very obvious.

c) *Value-Expectancy-Instrumentality Methodology*. In the course of a number of programme evaluations and cross-cultural studies, we have first asked people to say how important it was to them to undertake each of a large number of different sorts of activity and then how satisfied they were with their available opportunities to do each of the things they felt it was personally important for them to do. Thereafter they were asked to select the *most* important of their important sources of dissatisfaction and indicate what would

GRID 20.2. A MODEL OF COMPETENCE

Examples of Potentially Valued Styles of Behaviour

Examples of components of effective behaviour.	**Achievement**				**Affiliation**				**Power**		
	Doing things which have not been done before.	Inventing things.	Doing things more efficiently than they have been done before.	Developing new formal scientific theories.	Providing support and facilitation for someone concerned with achievement.	Establishing warm, convivial relationships with others.	Ensuring that a group works together without conflict.	Establishing effective group discussion procedures.	Ensuring that group members share their knowledge so that good decisions can be taken.	Articulating group goals and releasing the energies of others in pursuit of them.	Setting up domino-like chains of influence to get people to do as one wishes without having to contact them directly.
Cognitive											
Thinking (by opening one's mind to experience, dreaming, and using other sub-conscious processes) about what is to be achieved and how it is to be achieved.											
Anticipating obstacles to achievement and taking steps to avoid them.											
Analysing the effects of one's actions to discover what they have to tell one about the nature of the situation one is dealing with.											
Making one's value conflicts explicit and trying to resolve them.											
Consequence anticipated: *Personal*: e.g. "I know there will be difficulties, but I know from my previous experience that I can find ways round them. *Personal normative beliefs*: e.g. "I would have to be more devious and manipulative than I would like to be to do that." *Social normative beliefs*: e.g. "My friends would approve if I did that": "It would not be appropriate for someone in my position to do that."											
Affective											
Turning one's emotions into the task: Admitting and harnessing feelings of delight and frustration: using the unpleasantness of tasks one needs to complete as an incentive to get on with them rather than as an excuse to avoid them.											
Anticipating the delights of success and the misery of failure.											
Using one's feelings to initiate action, monitor its effects, and change one's behaviour.											
Conative											
Putting in extra effort to reduce the likelihood of failure.											
Persisting over a long period, alternatively striving and relaxing.											
Habits and experience											
Confidence, based on experience, that one can adventure into the unknown and overcome difficulties, (This involves knowledge that one will be able to do it plus a stockpile of relevant habits).											
A range of appropriate routineised, but flexibly contingent behaviours, each triggered by cues which one may not be able to articulate and which may be imperceptible to others.											
Experience of the satisfactions which have come from having accomplished similar tasks in the past.											

happen if they were to try to do something about the problem. The potential consequences studied were drawn from the range indicated by Fishbein (1967). They therefore included questions about whether they would be able to gain the satisfactions which they personally wanted, whether they would be able to live up to their personal – moral – self-images, and what reactions they anticipated from reference groups.

The results obtained from programme evaluations and organisational surveys conducted in this way have been extremely revealing.

One study (Raven, 1980) generated numerous new insights into parents' and teachers' child rearing behaviour. Mothers tend to create individualised, competency-oriented, developmental programmes for their children. But, although some teachers would like to do this, they do not know enough about each of their pupils to so. And the attempt to do so often confronts them with a host of moral dilemmas: Should they, for example, encourage independence and question-asking among children who live in dangerous environments and have parents who cannot manage independent children who are liable to question commands? In the course of the study a whole new set of issues bearing on parents' and teachers' competence in child rearing – and the tools they would need if they are to behave more competently – came to light. More specifically, exploration of teachers' and parents' competence to pursue their own lives and do their jobs effectively showed that many were in no position to provide appropriate role models for children. It follows that, if one wishes to facilitate the growth of competence in children, one extremely important starting point is by enhancing the competence of *their caregivers* to do the things *they* want and need to do. It also turned out that it was mothers' lack of confidence in their own competence as mothers which led them to hand their children over to other caretakers, but, paradoxically, those professionals were in no position to nurture the children's most important competencies.

In another study (Raven, Johnstone & Varley, 1985; Raven & Varley, 1984) the methodology was used to assess the effect that different teachers had on children's awareness of their motives, their values and priorities, and their competence to undertake activities they cared about effectively. It emerged that, contrary to the claim that schools make no difference, teachers had dramatic, and markedly different, effects on pupils' values and the consequences they expected if they were to set about tackling

problems they cared about. Most importantly, it emerged that previous evaluations of interdisciplinary, enquiry-oriented, project-based education had been entirely – and damagingly – misleading. Properly organised, project-based education has dramatic, positive, effects on children's confidence and competence. There is, however, a fundamental problem which prevents generalisation of the work. This is that there are no good tools to help teachers identify each child's motives, create individualised, competency-oriented, developmental programmes, and monitor each child's growth.

In a third study (Raven, 1984; Graham & Raven, 1987) it was found that, as in McClelland's (1961) work, there are dramatic differences between the pre-occupations of people who live in different societies and their willingness to do the things that are necessary – that is to say, their competence – to translate those values into effect. As far as can be judged, these differences are directly related to the kind of society which develops.

What these studies show is that the application of value-expectancy-instrumentality methodology guided by the framework for thinking about competence and test validation developed above does yield information which is more revealing, more valid, more comprehensive, and therefore more ethical, than that which would have been obtained had the studies been conducted only with tests of the kind which the *Joint Committee on the Evaluation of Educational Programs and Policies* (Stufflebeam, 1981) enjoin us to use – namely tests which have been shown to be reliable and valid in the conventional sense.

The application of value-expectancy-instrumentality methodology to the assessment of individual competence has proved more cumbersome than its application in programme evaluation. Nevertheless, computerised tools in this area are now available (Raven & Sime, 1994).

Summary and Conclusion

Three sets of conclusions – at different levels – emerge from what has been said in this article.

The first is that Spearman appears to have been right to emphasise the distinctive psychological nature of eductive and reproductive ability and to argue these abilities have different genetic and environmental determinants and different consequences for people's lives. These aspects of "intelligence" emerge as being among the most important variables psychology – whether pure or applied, whether "cognitive", "educational", or "occupational" – deals with.

The underlying *reasons* for – i.e. the interpretation to be placed upon – this now well established network of relationships, is, however, seriously open to question. At this point in time it is possible to attribute the entire observed pattern of relationships to variation, not in *cognitive* ability, but in motivational predispositions.

Even setting that disturbing thought on one side, the material reviewed shows that there is, at the very least, an urgent need to reconsider the way we think about and assess problem-solving ability, intelligence, and competence. At a minimum, we need more appropriate ways of conceptualising and assessing "cognitive ability". More basically, and more importantly, it is vital to broaden our framework for thinking about competence and social functioning so that we can situate our assessments of eductive ability in the context of assessments of (i) motives in the service of which eductive ability (as a component of competence) is applied, and (ii) other components of competence. Without such developments, our assessments of both individuals and educational programmes appear to be unethical. This is because they are insufficiently comprehensive and , as a result, lead to practices which are not in the best long-term interests of the individuals or programmes being evaluated – and therefore not in the long-term interests of society. The psychometric model which is required to come to terms with this problem differs markedly from that which has been pre-eminent in the past. A two-stage measurement process must be envisaged. We must first identify people's motives or valued styles of behaviour and then ask which of a range of cognitive, affective, and conative competencies they bring to bear in their efforts to undertake the activities they care about. A number of ways in which this model has been operationalised have been presented, but a great deal of further development work is required.

But what has been said also appears to have implications at a quite different level. We need to fundamentally reconsider the way in which seek to establish the validity of tests. On the one hand, the validity of the *criteria* as indices of the underlying construct we are seeking to assess is a much more serious problem than it has usually been taken to be. Behaviour is a poor index of psychological constructs since what people do depends on very many things, some arising from personal and value conflicts, some from environmental constraints. To find out what people are doing one needs somehow to get inside their heads. Setting them an alternative task – "performance assessment" – does not solve the problem. Thereafter one needs to somehow to examine the way in which motivational dispositions, eductive ability, and other components of competence contribute to that performance. What is required is fundamentally a conceptual, rather than a statistical, exercise … although path analysis certainly has a role to play.

References

Barrett, G.V. and Depinet, R.L. (1991). A reconsideration of testing for competence rather than intelligence. *American Psychologist*, *46(10)*, 1012-1024.

Berg, I. (1973). *Education and Jobs: The Great Training Robbery*. London: Penguin Books.

Bouchard, T.J. (1991). A twice told tale: Twins reared apart. In W.Grove and D.Ciccehetti (Eds.). *Thinking Clearly about Psychology: Essays in Honor of Paul Everett Meehl: Personality and Psychopathology*, (Volume 2).

Bouchard, T.J. and McGue, M. (1990). Genetic and rearing environmental influences on adult personality: An analysis of adopted twins reared apart. *Journal of Personality*, March.

Brown, J.S., Collins, A. and Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, January-February, 32-42.

Burgess, T. and Adams, E. (1980). *Outcomes of Education*. London: MacMillan Education.

Burgess, T. and Adams, E. (1986). *Records of Achievement at 16*. Windsor, Berks: NFER-Nelson.

Cattell, R.B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1-22.

Carpenter, P.A., Just, M.A. and Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97(3)*, 404-431.

Chomsky, N. (1987). *The Chomsky Reader*. London: Serpent's Tail.

Court, J., & Raven, C. J. (2001). *A Researcher's Bibliography for Raven's Progressive Matrices and Mill Hill Vocabulary Scales*. Obtainable in hard copy and disk format from Susan Middleton, Harcourt Assessment, 19500 Bulverde Rd., San Antonio, Texas 78259, USA. <susan_middleton@harcourt.com>

Court, J.H. and Raven, J. (1995). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.7, Research and References, Summaries of Normative, Reliability, and Validity Studies and References to all Sections*. Oxford, England, San Antonio, TX: Harcourt Assessment.

Crichton Royal, Dumfries. (1957). *A Comparative Study of Two Psychological Tests*. Department of Psychological Research. Unpublished manuscript.

Day, P. and Klein, R. (1987). *Accountabilities: Five Public Services*. London: Tavistock Publications.

Dalziel, M.M. and Schoonover, S.C. (1988). *Changing Ways*. New York: American Management Association.

de Leeuw, J. and Meester, A.C. (1984). Over het intelligente – onderzoek bij de militaire keuringen vanaf 1925 tot henden. [Intelligence – as tested at selections for the military service from 1925 to the present]. *Mens en Maatschappij,* 59, 5-26.

Deary, I.J. (1993). Inspection time and WAIS-R IQ subtypes: A confirmatory factor analysis study. *Intelligence, 17*, 223-236.

Deary, I.J. (1995). Auditory inspection time and intelligence: What is the direction of causation? *Development Psychology, 31*, 237-250.

Deming, W.E. (1993). *The New Economics for Industry, Government, and Education*. Cambridge, MA: Massachusetts Institute of Technology.

Eysenck, H.J. (1953). *Uses and Abuses of Psychology*. Harmondsworth, Mddx: Penguin Books.

Feuerstein, R., Klein, P., and Tannenbaum, A. (Eds.). (1990). Mediated learning experience: Theoretical, psychosocial, and educational implications. *Proceedings of the First International Conference on Mediated Learning Experience*. Tel Aviv: Freund.

Fishbein, M. (Ed.). (1967). *Readings in Attitude Theory and Measurement*. New York: Wiley.

Fivars, G. and Gosnell, D. (1966). *Nursing Evaluation*: *The Problem and the Process*. Pittsburg PA: Westinghouse Learning Corp.

Flanagan, J.C. (1949). Critical Requirements: A new approach to employee evaluation. *Personnel Psychology, 2*, 419-425.

Flanagan, J.C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.

Flynn, J.R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171-191.

Foulds, G.A. and Raven, J.C. (1948). Intellectual ability and occupational grade. *Occupational Psychology, 22*, 197-203.

Fraser-Roberts, J.A. (1943). *Further Observations on the Efficiency of the Progressive Matrices Test*. London: War Cabinet Expert Committee Report.

Gallimore, R. (1985). *The Accommodation of Instruction to Cultural Differences*. Los Angeles: University of California, Department of Psychiatry.

Gardner, H. (1985). *The Mind's New Science*. New York: Basic Books.

Gardner, H. (1987). Developing the spectrum of human intelligence. *Harvard Education Review, 57*, 187-193.

Gardner, H. (1991). Assessment in context: The alternative to standardized testing. In B.R. Gifford and M.C. O'Connor (Eds.), *Changing Assessments*: *Alternative Views of Aptitude, Achievement and Instruction*. New York: Kluwer Publishers.

Ghiselli, E.E. (1966). *The Validity of Occupational Aptitude Tests*. New York: Wiley.

Graham, M.A. and Raven, J. (1987). *International Shifts in the Workplace – are we becoming an "Old West" in the Next Century?* Provo: BYU, Department of Organizational Behavior.

Greeno, J.G. (1989). A perspective on thinking. *American Psychologist, 44(2)*, 134-141.

Guilford, J.P. (1977). *Way Beyond the IQ*. New York: Creative Education Foundation and Creative Synergetic Associates.

Hatch, T.C. and Gardner, H. (1986). From testing intelligence to assessing competencies: A pluralistic view of the intellect. *Roeper Review, 8*, 147-150.

Hatch, T.C. and Gardner, H. (1990). If Binet had looked beyond the classroom: The assessment of multiple intelligences. *International Journal of Educational Research*, 415-429.

Hogan, R. (1990). Unmasking incompetent managers. *Insight*, May 21, 42-44.

Hogan, R. (1991). *An Alternative Model of Managerial Effectiveness*. Mimeo: Tulsa, OK: Institute of Behavioral Sciences.

Hogan, R., Raskin, R. and Fazzini, D. (1990). The dark side of charisma. In K.E. Clark and M.B. Clark (Eds.), *Measures of Leadership*. West Orange, NJ: Leadership Library of America.

Holland, J.L. and Richards, J.M. (1965). Academic and non-academic accomplishments. *Journal of Educational Psychology, 56*, 165-175.

Hope, K. (1984). *As Others See Us: Schooling and Social Mobility in Scotland and the United States*. New York: Cambridge University Press.

Horn, J.L. (1968). Organization of abilities and the development of intelligence. *Psychological Review, 72*, 242-259.

Horn, J.L. (1994). Theory of fluid and crystallized intelligence. In R.J. Sternberg (Ed.), *Encyclopaedia of Human Intelligence* (443-451). New York: Macmillan.

House, E.R. (1991). Realism in research. *Educational Researcher, 20*, 2-9.

Huff, S., Lake, D. and Schaalman, M.L. (1982). *Principal Differences: Excellence in School Leadership and Management*. Boston: McBer and Co.

Hunter, J.E. and Hunter R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Ingleton, C.C.P. (1990). Personal communication, based on the study reported by C.C.P. Ingleton, *The Use of Tests in Graduate Selection* (61-76). In K.M. Miller (Ed.), (1975), *Psychological Testing in Personnel Assessment*. Epping, UK: Gower Press.

Jackson, P.W. (1986). *The Practice of Teaching*. New York: Teachers College Press.

Jacobs, P.I. and Vandeventer, M. (1968). Progressive Matrices: An experimental, developmental, nonfactorial analysis. *Perceptual and Motor Skills, 27*, 759-766.

Jaques, E. (1976). *A General Theory of Bureaucracy*. London: Heinemann.

Jaques, E. (1989). *Requisite Organization*. Arlington, VA: Cason Hall and Co.

Jensen, A.R., Larson, G.E. and Paul, S.M. (1988). Psychometric ***g*** and mental processing speed on a semanitic verification test. *Personality and Individual Differences, 9(2)*, 243-255.

Kanter, R.M. (1985). *The Change Masters: Corporate Entrepreneurs at Work*. Hemel Hempstead: Unwin Paperbacks.

Kinsey, A.C. (1948). *Sexual Behavior in the Human Male*. New York: Saunders.

Klemp, G.O., Munger, M.T. and Spencer, L.M. (1977). *An Analysis of Leadership and Management Competencies of Commissioned and Non-Commissioned Naval Officers in the Pacific and Atlantic Fleets*. Boston: McBer.

Kohn, M.L. (1969/77). *Class and Conformity: A Study in Values (Second Edition)*. Chicago IL: Chicago University Press. (1st edition: Dorsey Press.)

Kohn, M.L. and Schooler, C. (1978). The reciprocal effects of the substantive complexity of work and intellectual flexibility: A longitudinal assessment. *American Journal of Sociology, 84*, 24-52.

Kohn, M.L. and Schooler, C. (1982). Job conditions and personality: A longitudinal assessment of their reciprocal effects. *American Journal of Sociology, 87*, 1257-86.

Kohn, M.L., Slomczynski, K.M. and Schoenbach, C. (1986). Social stratification and the transmission of values in the family: A cross-national assessment. *Sociological Forum*, 1.

Lempert, W. (1986). Sozialisation und Personlicckeitsbildung in beruflichen Schulen, dargestellt am Beispiel der Entwicklung moralischer Orientierung. *Die berufsbildende Schule, 38*, 148-160.

Lempert, W., Hoff, E.H. and Lappe, L. (1990). *Occupational Biography and Personality Development: A Longitudinal Study of Skilled Industrial Workers*. Berlin: Max Planck Institute for Human Development and Education.

McClelland, D.C. (1951). *Personality*. New York: Sloane, Dryden, Holt. Reprinted by Irvington Publishers, New Jersey.

McClelland, D.C. (1961). *The Achieving Society*. New York: Van Nostrand.

McClelland, D.C. (1973). Testing for competence rather than for "intelligence". *American Psychologist, 28*, 1-14.

McClelland, D.C. (1978). *Guide to Behavioral Event Interviewing*. Boston: McBer.

McClelland, D.C, Atkinson, J.W., Clark, R.A. and Lowell, E.L. (1958). A scoring manual for the achievement motive. In J.W. Atkinson (Ed.), *Motives in Fantasy, Action and Society*. New York: Van Nostrand.

McClelland, D.C. and Dailey, C. (1973). *Evaluating New Methods of Measuring the Qualities Needed in Superior Foreign Service Workers*. Boston: McBer and Co.

McClelland, D.C. and Dailey, C. (1974). *Professional Competencies of Human Service Workers*. Boston: McBer and Co.

MacPherson, J.S. (1958). *Eleven Year Olds Grow Up*. London: University of London Press.

Maistriaux, R. (1959). *L'Intelligence et le Caractere*. Paris, France: Presses Universitaires de France.

Matarazzo, J.D. (1990). Psychological assessment versus psychological testing. *American Psychologist, 45*, 999-1017.

Maxwell, J.N. (1961). *The Level and Trend of National Intelligence: The Contribution of the Scottish Mental Surveys*. London: University of London Press.

Maxwell, J.N. (1969). *Sixteen Years On*. Edinburgh: Scottish Council for Research in Education.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18(2)*, 5-11.

Moreland, K.L., Eyde, L.D., Robertson, G.J., Primoff, E.S. and Most, R.B. (1995). Assessment of test user qualifications. *American Psychologist, 50*, 14-23.

Nuttgens, P. (1988). *What Should We Teach and How Should We Teach It*? Aldershot: Wildwood House.

Ogbu, J.U. (1992). Understanding cultural diversity and learning. *Educational Researcher, 21*, 5-14.

Rasch, G. (1947). In G. Rasch (1980), *Probabalistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.

Raven, J. (1976). *Pupil Motivation and Values*. Dublin: Irish Association for Curriculum Development.

Raven, J. (1977). *Education, Values and Society: The Objectives of Education and the Nature and Development of Competence*. London, England, H.K. Lewis. Now available from the author at 30 great King St., Edinburgh EH3 6QH, Scotland..

Raven, J. (1980). *Parents, Teachers and Children: An Evaluation of an Educational Home Visiting Programme*. Edinburgh: Scottish Council for Research in

Education. Distributed in North America by the Ontario Institute for Studies in Education, Toronto.

Raven, J. (1984). *Competence in Modern Society: Its Identification, Development and Release*. Unionville, New York: Trillium Press.

Raven, J. (1987). Values, diversity and cognitive development. *Teachers College Record, 89*, 21-38.

Raven, J. (1988). Toward measures of high-level competencies: A re-examination of McClelland's distinction between needs and values. *Human Relations, 41*, 281-294.

Raven, J. (1989a). Parents, education and schooling. In C. Desforges (Ed.), *British Journal of Educational Psychology, Monograph Series No.4, Special Issue on Early Childhood Education* (47-67).

Raven, J. (1989b). The Raven Progressive Matrices: A review of national norming studies and ethnic and socio-economic variation within the United States. *Journal of Educational Measurement, 26*, 1-16.

Raven, J. (1991). *The Tragic Illusion: Educational Testing*. New York: Trillium Press.

Raven, J. (1994). *Managing Education for Effective Schooling: The Most Important Problem is to Come to Terms with Values*. New York: Trillium Press.

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1-48.

Raven, J. and Dolphin, T. (1978). *The Consequences of Behaving: The Ability of Irish Organisations to Tap Know-How, Initiative, Leadership and Goodwill*. Edinburgh: The Competency Motivation Project.

Raven, J., Hannon, B., Handy, R., Benson, C. and Henry, E.A. (1975). *A Survey of Attitudes of Post Primary Teachers and Pupils, Volume 2: Pupils' Perceptions of Educational Objectives and their Reactions to School and School Subjects*. Dublin: Irish Association for Curriculum Development.

Raven, J., Johnstone, J. and Varley, T. (1985). *Opening the Primary Classroom*. Edinburgh: Scottish Council for Research in Education.

Raven, J., Raven, J. C., & Court, J. H. (1998, updated 2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1: General Overview*. San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3: The Standard Progressive Matrices,Iincluding the Parallel and Plus Versions*. San Antonio, TX: Harcourt Assessment.

Raven, J., Ritchie, J. and Baxter, D. (1971). Factor analysis and cluster analysis: Their value and stability in social survey research. *Economic and Social Review*, 367-391.

Raven, J. and Sime, J. (1994). *Computerised Edinburgh Questionnaires*. Edinburgh, The Competency Motivation Project.

Raven, J., & Stephenson, J. (Eds.). (2001). *Competence in the Learning Society.* New York: Peter Lang.

Raven, J. and Varley, T. (1984). Some classrooms and their effects: A study of the feasibility of measuring some of the broader outcomes of education. *Collected Original Resources in Education, 8(1)*, F4 G6.

Raven, J.C. (1939). The RECI series of perceptual tests: An experimental survey. *British Journal of Medical Psychology, XVIII*, Part 1, 16-34.

Ree, M.J., Earles, J.A. and Teachout, M.S. (1994). Predicting job performance: Not much more than ***g***. *Journal of Applied Psychology, 79*, 518-524.

Richardson, K. (1991). Reasoning with Raven – in and out of context. *British Journal of Educational Psychology, 61*, 129-138.

Roberts, E.B. (1968). A basic study of innovators: How to keep and capitalize on their talents. *Research Management, XI*, 249-266.

Rogers, E.M. (1962/83). *Diffusion of Innovations*. New York: Free Press.

Scottish Council for Research in Education (1933). *The Intelligence of Scottish Children*. London: University of London Press.

Scottish Council for Research in Education (1949). *The Trend of Scottish Intelligence*. London: University of London Press.

Scottish Council for Research in Education (1953). *Social Implications of the 1947 Scottish Mental Survey*. London: University of London Press.

Sigel, I.E. (1986). Early social experience and the development of representational competence. In W. Fowler (Ed.), *Early Experience and the Development of Competence. New Directions for Child Development, No.32*. San Francisco: Jossey-Bass.

Spearman, C. (1927). *The Abilities of Man*. London, England: MacMillan.

Spencer, L.M. and Spencer, S.M. (1993). *Competence at Work*. New York: Wiley.

Sperry, R. (1983). *Science and Moral Priority: Merging Mind, Brain, and Human Values*. Oxford: Blackwell.

Stansbury, D. (1980). The record of personal experience. In T. Burgess and E. Adams, *Outcomes of Education*. Basingstoke: MacMillan Education.

Sternberg, R.J., Wagner, R.K. and Okagaki, L. (1986). Practical intelligence: The nature and role of tacit knowledge in work and school. In R.J. Sternberg (Ed.), *Practical Intelligence*. New York: Cambridge University Press.

Stufflebeam Joint Committee on Standards for Educational Evaluation (1981). *Standards for Evaluations of Educational Programs, Projects and Materials*. New York: McGraw Hill.

Styles, I. (1999). The study of intelligence – the interplay between theory and measurement. In M. Anderson (Ed.), *The Development of Intelligence*. Sussex, England: Psychology Press.

Taylor, C.W. (1971). *All of Our Children are Educationally Underprivileged*. Salt Lake City: Department of Psychology, University of Utah.

Taylor, C.W. (1976). *Talent Ignition Guide*. Salt Lake City: University of Utah and Bellvista Public School.

Taylor, C.W. and Barron, F. (Eds.). (1963). *Scientific Creativity: Its Recognition and Development*. New York: Wiley.

Taylor, C.W., Smith, W.R. and Ghiselin, B. (1963). The creative and other contributions of one sample of research scientists. In C.W. Taylor and F. Barron (Eds.), *Scientific Creativity*. New York: Wiley.

Tellegen, A., Lykken, D.T., Bouchard, T.J., Wilcox, K.J., Segal, N.L. and Rich, S. (1988). Personality similarity in twins reared apart and together. *Journal of Personality and Social Psychology, 54(6)*, 1031-1039.

Tharp, R.G., Jordan, C., Speidel, G.E., Au, K.H.P., Klein, T.W., Calkins, R.P., Sloat, K.C.M. and Gallimore, R. (1984). Product and process in applied developmental research: Education and the children of a minority. In M.E. Lamb, A.L. Brown and B. Rogoff (Eds.), *Advances in Developmental Psychology, Volume III* (91-144). Hillsdale, NJ: Lawrence Erlbaum.

Trevarthen, C. (1990). Growth and education of the hemispheres. In C. Trevarthen (Ed.), *Brain, Circuits and Functions of the Mind: Essays in Honor of Roger W. Sperry*. Cambridge, England: Cambridge University Press.

Trevarthen, C. (1992). The self born in inter-subjectivity: The psychology of infant communicating. In U. Neiser (Ed.), *Ecological and Interpersonal Knowledge of Self*. New York: Cambridge University Press.

Van Beinum, H. (1965). *The Morale of the Dublin Busman*. London: Tavistock Institute of Human Relations.

Vernon, P.A. (1989). *Speed of Information-Processing and Intelligence*. Norwood, NJ: Ablex.

Vernon, P.A. (1991). Studying intelligence the hard way. *Intelligence, 15*, 389-395.

Vernon, P.A. (1993). Intelligence and neural efficiency. In D.K. Detterman (Ed.), *Current Topics in Human Intelligence, Volume 3*. Norwood, NJ: Ablex.

Vernon, P.E. and Parry, J.B. (1949). *Personnel Selection in the British Forces*. London: University of London Press.

Vodegel-Matzen, L.B.L. (1994a). *Performance on Raven's Progressive Matrices*. Ph.D. Thesis, University of Amsterdam.

Vodegel-Matzen, L.B.L., van der Molen, M.W. and Dudink, A.C.M. (1994b). Error analysis of Raven test performance. *Personality and Individual Differences, 16(3)*, 433-445.

Waller, N.G., Kojetin, B.A., Bouchard, T.J., Lykken, D.T. and Tellegen, A. (1989). Genetic and environmental influences on religious interests, attitudes and values: A study of twins reared apart and together. *Psychological Science, 1(2)*, March, 138-142.

Vygotsky, L.S. (1981). The genesis of higher mental function. In J.V. Wertsch (Ed.), *The Concept of Activity in Society Psychology*. Annank, NH: Sharpe.

Winter, D.G. (1973). *The Power Motive*. New York: Free Press.

Wright, B.D. and Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

PART V

EMERGING APPLICATIONS

The chapters in this section look at emerging applications of the RPM in predicting such things as driver accidents and malingering in court proceedings.

The section should have included a chapter on the deployment of the RPM in facilitating or denying claims for payment for such things as remedial education or health care. Unfortunately, this chapter was not produced in a timely manner.

Chapter 21

Predicting Driver Behaviour

Joerg Prieler

Background

Psychological assessment of individuals who have been disqualified from driving and who now wish to renew their licences and those seeking employment as, for example, bus drivers has become a major industry in some countries. Yet the task of assessing and predicting driver behaviour and accident proneness has proved to be no easy matter and, as is shown in this chapter, yields some surprising results. The results reported here are fairly typical of those emerging from a number of validation studies conducted in different countries.

Population studied

The total group from which those involved in the present study were chosen consisted of 786 applicants for positions as bus drivers. Following training, they were assessed by their instructors and recommended, or otherwise, for employment.

In order to eliminate the effects of variation between instructors, the present study focused on 229 drivers who had been assessed by a single instructor. 125 of these were recommended for employment and 103 not recommended.

In addition, a study was made of the driver errors recorded by the instructors and efforts made to predict both instructor recommendation and driver errors from batteries of psychological tests.

Assessment of driver competence

The aspects of driver behaviour assessed by the instructor are given in Table 21.1.

Table 21.1. Components of Driver Behaviour Assessed by Instructor

Vehicle Check outside	Observation/blindspots
Vehicle Check inside	Positioning
Reversing	Turning
Steering	Spacing between vehicles
Clutch	Overtaking
Brakes	Intersection
Acceleration	Driving: downhill
Gears	Driving: uphill
Indicators	Attitude eg aggressive
Horn	Courtesy
Mirrors	Ability to talk and drive at same time.

Because there was too little variance on each of the variables (turning, use of horn, etc.) taken singly, an overall index of the total number of errors made by each driver was computed.

Psychological assessment

The tests were administered by computer using the Vienna Test System developed by G. Schuhfried, GmbH, Austria. The tests selected had been validated in previous studies conducted by that company.

The test battery consisted of:

- COG - Cognitrone (Concentration test), Testform: S9
- DT - Determination Test (Stress test), Testform: S5
- RT - Reaction Test, Testform: S3
- LVT - Visual Pursuit Test (Perception test) , Testform: S2
- ZBA - Time-Movement Anticipation, Testform: S3
- 16PF - 16-Personality Factor-Test (Cattell)
- SPM - Raven's Standard Progressive Matrices

Results

Predicting whether a Driver will be recommended or not.

The psychological test variables on which the drivers who were recommended for employment differed significantly from those not recommended are shown in Table 21.2.

From the results relating to the Concentration test (Cog), it would seem that recommended bus drivers show significantly more correct reactions and their reaction times are shorter, regardless of whether their reactions are correct or incorrect.

The time-movement anticipation (ZBA) test scores of the recommended drivers are also better; they can better anticipate where e.g. a car will be on the street some moments later.

On the Stress Test (DT), recommended bus drivers make significantly more correct reactions, and less incorrect, delayed and omitted reactions, and are generally faster.

The SPM scores of the recommended drivers are significantly higher than those who are rejected.

And, on the 16PF, the recommended bus drivers are significant more sociable, obtain higher abstract thinking scores, have a greater sense of duty, and generally think more carefully about things.

Predicting driver errors

As previously mentioned, the analysis was carried out using a single overall index of number of errors.

To simplify the analysis, the test scores from the psychological tests were first factor analysed. These analyses were run separately for the ability and personality tests. Because the initial results yielded some counter-intuitive results for the reaction time tests, these were dropped from the analysis. The results of the revised analyses are shown in table 21.3 and 21.4.

From Table 21.3, it appears that only Factor 6, whose main loading is on the SPM, is significantly able to predict driver errors. None of the personality factors predict the total driver error score.

These results are typical of those found in many traffic validation studies conducted by the author and his colleagues, and they also replicate the results of many wider test validation studies conducted since the Second World War and summarised in other chapters of this book. Time and again it emerges that, when some measure of general cognitive ability (of which the *Raven's Progressive Matrices* is the most cost-effective) is included in a battery, it mops up virtually all the predictive validity of the other tests. The results reported here strongly reinforce the conclusions

Table 21.2. Significant Differences Between Recommended and Not Recommended Drivers

Variable name	*Mean value: Recommended drivers*	*Mean value: Not recommended drivers*	*Significance*
Cog Sum "correct reactions"	437.72	372.43	p = .000
Cog Mean time "correct reactions" (sec)	.87	1.02	p =.001
Cog Mean time "incorrect reactions" (sec)	.84	1.02	p = .016
Zba Median direction deviation (total)	26.86	33.23	p = .028
Dt Median reaction time	.83	.92	p = .000
Dt On time	335.50	233.74	p = .000
Dt Delayed	127.93	166.44	p = .000
Dt Incorrect	62.55	103.71	p = .041
Dt Omitted	55.46	105.06	p = .000
SPM Raw score	35.56	29.67	p = .000
16PF_A: Sociability	5.28	4.77	p = .010
16PF_B: Abstract thinking	5.32	4.76	p = .028
16PF_G: Sense of duty	6.32	5.48	p = .000
16PF_N: Thinking carefully	4.90	4.50	p = .031

Table 21.3. Factor Analysis: Ability Tests Without Reaction Test

Rotated Components

	Component					
	1	2	3	4	5	6
Age	.032	.574	.048	-.008	-.023	-.075
cog Sum “Reactions” (right and wrong Reactions)	.934	-.229	.071	-.118	-.099	.025
cog Sum “right Reactions”	.930	-.235	-.090	-.113	-.103	.007
cog Sum “wrong Reactions”	.103	.018	.971	-.041	.015	.109
cog percent range “wrong Reactions”	-.072	.040	.974	-.009	.004	.046
cog Mean time “right Reactions” (sec)	-.938	.210	-.032	.130	.051	-.005
cog Mean time “wrong Reactions” (sec)	-.821	.172	-.031	.136	-.151	-.014
lvt Median time right answers (sec)	-.168	-.003	-.090	.932	.045	-.005
lvt score	.220	-.012	-.044	-.929	-.095	-.019
zba Median Deviation time	.023	.146	.011	.049	.658	-.032
zba Median Median direction	-.028	.170	-.032	.075	.770	.024
dt Median Reaction time (Modus Reaction)	-.201	.950	.000	.015	.104	.027
dt on time (Modus Reaction)	.227	-.929	-.033	.007	-.162	-.066
dt delayed (Modus Reaction)	-.200	.894	.019	.049	.006	.036
dt Incorrect (Modus Reaction)	.068	-.178	.235	-.220	.017	.403
dt Omitted (Modus Reaction)	-.197	.749	-.009	-.023	.278	-.072
SPM raw score	.341	.127	-.216	-.087	-.396	-.576
driving errors total	.108	-.066	-.178	.219	-.345	.579

Extraction method: Main components Rotation method: Varimax with Kaiser-Normalization

a. The Rotation is converged in 7 Iterations.

Table 21.4. Factor Analysis of 16 PF Test

	Component				
	1	2	3	4	5
Warmth (Reserved vs. Warm; Factor A)	-,182	,054	-,091	,769	,093
Reasoning (Concrete vs. Abstract; Factor B)	-,033	,704	,292	,139	-,232
Emotional Stability (Reactive vs. Emotionally Stable; Factor C)	-,488	,401	-,356	,109	,106
Dominance (Deferential vs. Dominant; Factor E)	,058	-,495	,574	-,072	-,132
Liveliness (Serious vs. Lively; Factor F)	,043	-,001	,136	,757	-,082
Rule-Consciousness (Expedient vs. Rule-Conscious; Factor G)	-,254	,596	-,123	-,150	,143
Social Boldness (Shy vs. Socially Bold; Factor H)	-,738	-,038	,091	,077	-,004
Sensitivity (Utilitarian vs. Sensitive; Factor I)	-,102	-,025	,061	-,004	,421
Vigilance (Trusting vs. Vigilant; Factor L)	,345	-,105	,471	,027	-,410
Abstractedness (Grounded vs. Abstracted; Factor M)	,121	-,070	,628	,143	,011
Privateness (Forthright vs. Private; Factor N)	,225	,481	-,113	,102	-,138
Apprehension (Self-Assured vs. Apprehensive; Factor O)	,712	,035	,058	-,057	-,060
Openness to Change (Traditional vs. Open to Change; Factor Q1)	-,151	,212	,713	-,118	,224
Self-Reliance (Group-Oriented vs. Self-Reliant; Factor Q2)	,521	-,090	,097	-,361	-,139
Perfectionism (Tolerates Disorder vs. Perfectionistic; Factor Q3)	-,546	,071	-,135	-,085	-,285
Tension (Relaxed vs. Tense; Factor Q4)	,549	-,487	-,014	,097	-,117
Driving errors total	,128	-,041	-,022	,030	,682

that stem from Carroll's 1993 book *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*: Attempts to "improve" the measurement of eductive and reproductive ability using "more basic" measures like "attention" and "reaction time", while greatly extending the time needed for testing (and improving the face validity of the battery), have added surprisingly little to either our scientific understanding or our professional competence.

Chapter 22

Detection of Adult Malingering on Raven's *Standard Progressive Matrices*: A Cross-Validation*

R. Kim McKinzey, Marvin H. Podd,
Mary Ann Krehbiel, and John Raven**

Abstract

A formula for detecting faked Raven's SPM profiles was cross-validated on 46 experimental malingerers and 381 people from the standardization sample. The formula yielded a cross-validated 26% false negative rate and a 5% false positive rate.

*****

David Faust (Faust, 1996; Faust, Ziskin, & Hiers, 1991) and Richard Rogers (Rogers, Harrell, & Liff, 1993) have documented the need for detection of neuropsychological malingering. To solve the problem, Gisli Gudjonsson and Harriet Shackleton (Gudjonsson & Shackleton, 1986) validated a formula using Raven's *Standard Progressive Matrices* (Raven, 1958). The formula has the distinct advantage of being usable on protocols given in the past, as it requires no special administration procedures.

The validation group consisted of 29 experimental malingerers (mean age 24 years), 56 normals (mean age 21-24 years), and 25 forensic patients (mean age 29 years) who had been referred for neuropsychological

* This article was previously published in the *British Journal of Clinical Psychology* (1999), 38, 435-439. © British Psychological Society and is reproduced with their permission.
** The authors would like to thank Victoria Campagna for her help in data collection.

evaluation. The formula compares the number of correct answers on the first 24 items against the number of correct answers in the last 24 items (the "rate of decay"), using a set of cut-off numbers derived from the expected, theoretical, rate of decay. Of the 29 malingerers, 5 (17%) were missed (i.e., were false negatives). Of the 81 honest patients and normals, 4 (5%) were incorrectly classified as faking (false positives).

However, malingering formulas have had a disappointing record of maintaining adequate accuracy on cross-validation (e.g., McKinzey & Russell, 1997a; McKinzey & Russell, 1997b). The formula was therefore replicated with a much larger, varied, sample.

Method, Results, and Discussion

Subjects

All 427 participants came from the community. Women comprised 56% of the sample. Age range was 17-91 (M = 44, SD = 17). The normal group consisted of 381 people drawn from the Dumfries standardization sample (Raven, Raven, & Court, 2000, updated 2004). Their ages ranged from 17 to 91, (M = 45, SD = 17). Women comprised 57%.

The standardization sample used socioeconomic status (SES) as measured by the Hall-Jones Scale of Occupational Prestige for Males. This scale ranks occupations on a 1-8 scale, with professionals ranked 1 and executives, skilled nonmanual workers, skilled and semi-skilled manual workers, and unskilled manual workers ranked progressively lower. Using this scale, the normal group's SES range is 1-8 (M = 5.24, SD = 1.84).

The malingering group consisted of 46 adults, age range 18 to 70, (M = 38, SD = 14). Their level of education (in years) was 4 to 20, (M = 14, SD = 3). Women comprised 52%.

Procedure

All Ss were given the Raven Standardized Progressive Matrices, using the 1998 norms (Raven et al., 1998). The test was administered according to standard instructions. The additional instructions given to the fakers were:

Pretend you have suffered head injuries in an accident caused by another person or persons. Assume you are involved in litigation to determine how much financial compensation you will obtain from the people responsible for the accident and/or from the insurance companies

involved. Imagine that your everyday functioning in and outside of school and/or work has been much worse since your accident, that your potential earning power has been substantially reduced, and that you deserve all the money that the courts will allow you. The results of this test will help determine how large your settlement will be, so fake the most severe disability that you can without making it obvious to the examiner that you are faking.

The Raven answers were applied to the formula (2A + B) - (D + 2E), where A, B, D, and E refer to the number of correct responses in each of the Raven subsets (Gudjonsson & Shackleton, 1986). The result of the formula is termed the "rate of decay", and is compared to the rate of decay by total score cutoffs (Table 22.1) validated in the original study. Since the current study has a larger, more varied, sample than in the original study, a gradual tapering was done to the originally abrupt cutoffs at the extremes of the total score range: The original study's cutoffs suggested that any perfect score (which gets a rate of decay of 0) must be a fake!

Results

Age, education, and total score were not significantly correlated with the formula's accuracy. The formula's classifications are presented in Table 22.2. The formula replicated Gudjonsson & Shackleton's false positive rate of 5%. The false negative rate changed from 17% to 26%, as expected for a cross-validation. If the base rate of malingering in a given population is assumed to be 10% (as it is in this sample), then a formula-based result of normal has a 97% chance of being correct, and a formula-based result of faked has a 63% chance of being correct. If the base rate is assumed to be truly unknown, and therefore assumed to be 50%, then a normal result has a 78% chance of being correct, and a faked result has a 93% chance of being correct. The hit rate would be 84%. While the formula will miss some people, a formula result of faked should be given considerable interpretive weight.

The false negative rate was not artificially elevated by the Ss' inability to fake the test. Other studies (e.g., Heaton, Smith, Lehman, & Vogt, 1978; McKinzey, Podd, Krehbiel, Mensch, & Trombka, 1997) have found that some Ss are unable to fake a given test sufficiently to produce abnormal results, a problem referred to as a "threat to external validity" (Rogers & Cruise, 1998). This problem can only be corrected when the

Table 22.1. Cutoff Values for Each Total Score

Total score	Cutoff	Total score	Cutoff	Total score	Cutoff	Total score	Cutoff	Total score	Cutoff
		13	9	25	11	37	10	49	6
2	1	14	9	26	11	38	8	50	6
3	2	15	9	27	11	39	8	51	6
4	3	16	9	28	12	40	8	52	6
5	4	17	9	29	12	41	8	53	2
6	5	18	10	30	12	42	8	54	2
7	6	19	10	31	12	43	7	55	2
8	7	20	10	32	12	44	7	56	2
9	7	21	10	33	10	45	7	57	0
10	7	22	10	34	10	46	7	58	0
11	7	23	11	35	10	47	7	59	-1
12	7	24	11	36	10	48	6	60	-1

Note. The rate of decay is calculated by comparing the number of correct answers in each subset according to the formula ((2*A)+B) - (D+(2*E)). The cutoff is determined by the total score. The Raven is considered invalid if the rate of decay is below the cutoff listed for each total score.

Table 22.2. Cross-Validation of the Formula: Classifications

	Formula Result		
	Raven faked: n (% of row)	Raven not faked: n (% of row)	Totals
Malingering group	34 (73.9)	12 (26.1)	46 (100)
Normal group	20 (5.25)	361 (94.75)	381 (100)
n	54	373	427

Note. Percentages are rounded. The chi-square statistic is highly significant: Chi-square = 175 (1), $p < .0001$.

test has a clear measure of abnormality, such as the Halstead Impairment Index. Such ineffectual faking attempts are of little consequence in interpretation, since the difference between the true and actual scores will be minimal. However, all of the 12 faking Ss missed by the formula (false negatives) yielded IQ scores in the 65-95 range, with seven of the 12 getting scores below 70. On the other hand, all but 4 of the 20 false positives coming from the standardization sample got IQ scores in the 98-135 range, with only one below 74. Any extremely low score should therefore be consistent with the available history and testing before being considered valid, even when the formula is negative.

Discussion

There are few methods of detecting faked IQ test results. Other, more accurate, methods are available to detect malingering of neuropsychological deficits: the Test Of Malingered Memory, a commercial product designed to detect neuropsychological malingering, is a stand-alone measure with a 2% false negative rate (Rees, Tombaugh, Gansler, & Moczynski, 1998). The Luria-Nebraska Neuropsychological Battery is a comprehensive neuropsychological test whose malingering formula has a 17% false negative rate (McKinzey et al., 1997). However, neither is an IQ test, and do not have the same place in a battery as the Raven.

There are many identifiable groups that were not included in the cross-validation. For example, there were no neurologically impaired patients, developmentally delayed participants, or forensic samples. The current subjects are all English-speaking, although the Raven is widely used with non-English speaking people. The faking formula should be cross-validated with such groups in future research, and interpretive caution employed until such research is done.

References

Faust, D. (1996). Neuropsychological (Brain Damage) Assessment. In J. Ziskin (Ed.), *Coping with Psychiatric and Psychological Testimony*, (5 ed., Vol. 2, pp. 916-1044). Los Angeles: Law and Psychology Press.

Faust, D., Ziskin, J., & Hiers, J. (1991). *Brain damage claims: coping with neuropsychological evidence*. Los Angeles: Law & Psychology Press.

Gudjonsson, G., & Shackleton, H. (1986). The pattern of scores on Raven's Matrices during faking bad and non-faking performance. *British Journal of Clinical Psychology, 25*, 35-41.

Heaton, R. K., Smith, H. H., Lehman, R. A., & Vogt, A. T. (1978). Prospects for faking believable deficits on neuropsychological testing. *Journal of Consulting & Clinical Psychology, 46* (5), 892-900.

McKinzey, R. K., Podd, M. H., Krehbiel, M. A., Mensch, A. J., & Trombka, C. C. (1997). Detection of malingering on the Luria-Nebraska Neuropsychological Battery: An initial and cross-validation. *Archives of Clinical Neuropsychology, 12* (5), 505-512.

McKinzey, R. K., & Russell, E. W. (1997a). Detection of malingering on the Halstead-Reitan Battery: A Cross-validation. *Archives of Clinical Neuropsychology, 12* (6), 585-590.

McKinzey, R. K., & Russell, E. W. (1997b). A partial cross-validation of a Halstead-Reitan Battery malingering formula. *Journal of Clinical and Experimental Neuropsychology, 19* (4), 484-488.

Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3: The Standard Progressive Matrices*. San Antonio, TX: The Psychological Corporation.

Raven, J. C. (1958). *The Standard Progressive Matrices*. London: H. K. Lewis. (An earlier version of this test was known as *Progressive Matrices (1938),* and also published by H. K. Lewis. The test was subsequently published by OPP Ltd. (Oxford) and now by Harcourt Assessment, San Antonio, TX.)

Rees, L. M., Tombaugh, T. N., Gansler, D. A., & Moczynski, N. P. (1998). Five validation experiments of the Test of Memory Malingering (TOMM). *Psychological Assessment, 10* (1), 10-20.

Rogers, R., & Cruise, K. R. (1998). Assessment of malingering with simulation designs: Threats to external validity. *Law and Human Behavior, 22* (3), 273-285.

Rogers, R., Harrell, E. H., & Liff, C. D. (1993). Feigning neuropsychological impairment: A critical review of methodological and clinical considerations. *Clinical Psychology Review, 13* (3), 255-274.

Chapter 23

Detection of Children's Malingering on Raven's *Standard Progressive Matrices**

R. Kim McKinzey, Jörg Prieler, and John Raven**

Abstract

A formula for detecting faked Raven's *Standard Progressive Matrices* profiles was cross-validated on 44 children and adolescents (ages 7-17). It yielded a false negative rate of 64%. However, a rule using three very easy items (i.e., any of A3, A4, or B1 missed) yielded a hit rate of 95%, with 5% false positive and negative rates. All but two of the participants were able to produce lower scores when asked to fake the test.

*****

David Faust (Faust, 1996; Faust, Ziskin, & Hiers, 1991) and Richard Rogers (Rogers, Harrell, & Liff, 1993) have documented the need for detection of neuropsychological malingering. To solve the problem, Gisli Gudjonsson and Harriet Shackleton (Gudjonsson & Shackleton, 1986) validated a formula using Raven's *Standard Progressive Matrices* (Raven, 1958). The formula has the distinct advantage of being usable on protocols given in the past, as it requires no special administration procedures. The formula was cross-validated (McKinzey, Podd, Krehbiel, & Raven, 1999) on 46 experimental malingerers and 381 people from the adult standardization sample (Raven, Raven, & Court, 2000). The formula yielded a cross-validated 26% false negative rate and a 5% false positive rate.

* This chapter was previously published in the *British Journal* of *Clinical Psychology (2003), 42*, 95-99. © British Psychological Society. Reproduced by permission.
** The authors would like to thank Brigitte Haider for her help in data collection.

However, the formula was validated and cross-validated on adults, and one does not have to be an adult to fake a psychological test. Faust has also documented the ability of normal adolescents (Faust, Hart, Guilmette, & Arkes, 1988) and children (Faust, Hart, & Guilmette, 1988) to produce abnormal neuropsychological test results that are believable and undetectable by the same methods that are useful in detecting cortically based neuropsychological deficits.

The formula was therefore applied to a sample of children and adolescents.

Method

Participants

All 44 participants came from schools in Vienna, Austria. All were white. Girls comprised 57% of the sample. Age range was 7-17 (M = 12.5, SD = 2.6), with 9 in the 7-10 age range, 27 in the 11-14 range, and 8 in the 15-17 range. Their total score range was 18-59, M = 43.95. Using the US norms (Raven, 2000), their percentile range was 24-99, M = 65.9.

Procedure

Parents signed consent forms. All participants were given Raven's *Standard Progressive Matrices*. The test was administered according to standard instructions (Raven et al., 2000). Then, the same participants were asked to take the test again, this time with the instructions (in German): *We know that some people don't try their best on this test. We'd like to find a way to catch them. To help us, please do as badly on this test as you can, without getting caught*.

The Raven answers were applied to the formula (2A + B) - (D + 2E), where A, B, D, and E refer to the number of correct responses in each of the Raven subsets (Gudjonsson & Shackleton, 1986; McKinzey et al., 1999). The result of the formula is termed the "rate of decay", and is compared to the rate of decay by total score cutoffs (see McKinzey et al., 1999, for details).

Results

In the faking condition, the participants produced total scores (range = 2-59, M = 9.6) and percentiles (range = 1-99, M = 6.3) substantially below that of the normal condition, with a mean difference between the two conditions of 34 ($t = -14.003$; $p < .0001$) and mean intrasubject total

score differences of 34 (SD = 15.36, range 1-57). Not surprisingly, using the formula cutoffs validated for adults proved highly inaccurate, with a false positive rate of 7% and a false negative rate of 64%.

The inaccuracy rates were minimally artificially elevated by the participants inability to fake the test. Other studies (e.g., Heaton, Smith, Lehman, & Vogt, 1978; McKinzey, Podd, Krehbiel, Mensch, & Trombka, 1997) have found that some participants are unable to fake a given test sufficiently to produce abnormal results, a problem referred to as a "threat to external validity" (Rogers & Cruise, 1998). Such ineffectual faking attempts are of little consequence in interpretation, since the difference between the true and actual scores will be minimal. However, one of the participants had no difference between the two conditions, and another did one item better! When these two participants were eliminated, the false negative rate dropped by only 2 percentage points. Changing the cutoff points was the obvious next step, but visual inspection revealed a far more obvious (and accurate) method of detecting the participants in the faking condition.

According to the standard instructions, the testee must agree to the correct answers to the first two items. The third item is very easy, and only one of the participants (an 8 year old girl who produced an average Raven IQ score) answered it incorrectly. When these same participants malingered on the test, item A3 was overwhelmingly answered incorrectly. Table 23.1 presents the classification results of the simple rule that a testee getting item 3 incorrect is showing insufficient effort. The hit rate of the rule is 90%, with a false positive rate of 2% and false negative rate of 17%. Combining the item 3 rule and the rate of decay formula decreased the hit rate by 1%.

Using analyses of item difficulty (Raven et al., 2000) as a guide, items A4 and B1 were similarly identified as increasing the hit rate. Table 23.2 presents the classification results, removing the two unsuccessful faking participants. Applying the rule that missing any of items A3, A4, or B1 (all extremely easy items) should be interpreted as showing insufficient effort produces a 95% hit rate, with equal false positive and negative rates of 5%. The two false positives were the 8 year old girl (normal condition total score = 21, 44th percentile) and an 11 year old girl (normal condition total score = 35, 34th percentile). The two false negatives were a 12 year old girl (faking condition total score = 19, down from 34) and a 14 year old girl (faking condition total score = 34, down from 53).

Table 23.1. Item 3 Rule Results

	Formula result		
	Raven faked: *n* (% of row)	Raven not faked: *n* (% of row)	Totals
Malingering condition	35 (83)	7 (17)	42 (100)
Normal condition	1 (2)	41(98)	42 (100)
n	36	48	84

Note. Percentages are rounded. The chi-square statistic is highly significant (Chi-square = 58.37, p <.0001; Fisher's Exact p < .0001). The table does not include the two participants who did not produce malingered scores.

Table 23.2. Item A3 or A4 or B1 Rule Results

	Formula result		
	Raven faked: *n* (% of row)	Raven not faked: *n* (% of row)	Totals
Malingering condition	40 (95)	2 (5)	42 (100)
Normal condition	2 (5)	40 (95)	42 (100)
n	42	42	84

Note. Percentages are rounded. The chi-square statistic is highly significant (Chi-square = 68.76, p <.0001; Fisher's Exact p < .0001). The table does not include the two participants who did not produce malingered scores.

Discussion

There are few methods of detecting faked IQ test results. Other cross-validated methods are available to detect malingering of neuropsychological deficits: for example, the *Test Of Malingered Memory* (TOMM), a commercial product designed to detect neuropsychological malingering, is a stand-alone measure with a 5% false negative rate (Rees, Tombaugh, Gansler, & Moczynski, 1998). The Luria-Nebraska Neuropsychological Battery (LNNB) is a comprehensive neuropsychological test whose within-test malingering formula has a 17% false negative rate (McKinzey et al., 1997). However, neither is an IQ test, and do not have the same place in a battery as the Raven. More importantly, no test, including the TOMM and LNNB, has published malingering measures validated on children or adolescents (as of this writing).

There are many identifiable groups that were not included in this study. For example, there were no neurologically impaired patients, people with tested IQs below 70, or forensic samples. The current subjects are all Austrian, although the Raven is widely used with people speaking a wide range of languages. The faking formula should be cross-validated with such groups in future research, and appropriate interpretive caution employed until such research is done.

References

Faust, D. (1996). Neuropsychological (brain damage) assessment. In J. Ziskin (Ed.), *Coping with Psychiatric and Psychological Testimony* (5 ed., Vol. 2, pp. 916-1044). Los Angeles: Law and Psychology Press.

Faust, D., Hart, K. J., & Guilmette, T. J. (1988). Pediatric malingering: The capacity of children to fake believable deficits on neuropsychological testing. *Journal of Consulting & Clinical Psychology, 56(4)*, 578-582.

Faust, D., Hart, K. J., Guilmette, T. J., & Arkes, H. R. (1988). Neuropsychologists' capacity to detect adolescent malingerers. *Professional Psychology: Research & Practice, 19(5)*, 508-515.

Faust, D., Ziskin, J., & Hiers, J. (1991). *Brain damage claims: coping with neuropsychological evidence*. Los Angeles: Law & Psychology Press.

Gudjonsson, G., & Shackleton, H. (1986). The pattern of scores on Raven's Matrices during "faking bad" and "non-faking" performance. *British Journal of Clinical Psychology, 25*, 35-41.

Heaton, R. K., Smith, H. H., Lehman, R. A., & Vogt, A. T. (1978). Prospects for faking believable deficits on neuropsychological testing. *Journal of Consulting & Clinical Psychology, 46(5)*, 892-900.

McKinzey, R. K., Podd, M. H., Krehbiel, M. A., Mensch, A. J., & Trombka, C. C. (1997). Detection of malingering on the Luria-Nebraska Neuropsychological Battery: An initial and cross-validation. *Archives of Clinical Neuropsychology, 12(5)*, 505-512.

McKinzey, R. K., Podd, M. H., Krehbiel, M. A., & Raven, J. (1999). Detection of Malingering on the Raven Progressive Matrices: A Cross-validation. *British Journal of Clinical Psychology, 38(3)*, 435-439.

Raven, J. (2000). Manual for Raven's Progressive Matrices and Vocabulary Scales. Research Supplement No.3 (Second Edition): A Compendium of International and North American Normative and Validity Studies Together with a Review of the Use of the RPM in Neuropsychological Assessment by Court, Drebing, & Hughes. San Antonio, TX: Harcourt Assessment.

Raven, J., Raven, J. C., & Court, J. H. (2000, updated 2004). Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3: The Standard Progressive Matrices. San Antonio, TX: Harcourt Assessment.

Raven, J. C. (1958). *The Standard Progressive Matrices*. London: H. K. Lewis. (An earlier version of this test was known as *Progressive Matrices (1938),* and also

published by H. K. Lewis. The test was subsequently published by OPP Ltd. (Oxford) and now by Harcourt Assessment, San Antonio, TX.)

Rees, L. M., Tombaugh, T. N., Gansler, D. A., & Moczynski, N. P. (1998). Five validation experiments of the Test of Memory Malingering (TOMM). *Psychological Assessment, 10(1)*, 10-20.

Rogers, R., & Cruise, K. R. (1998). Assessment of malingering with simulation designs: Threats to external validity. *Law and Human Behavior, 22(3)*, 273-285.

Rogers, R., Harrell, E. H., & Liff, C. D. (1993). Feigning neuropsychological impairment: A critical review of methodological and clinical considerations. *Clinical Psychology Review, 13(3)*, 255-274.

PART VI

SOME OUTSTANDING ETHICAL ISSUES

Although some extremely serious ethical issues were raised in the earlier chapter on "Intelligence, Engineered Invisibility, and the Destruction of Life on Earth", we now return to the question of the ethical application of tests – and the role of Psychologists in relation to them.

In the first chapter, "Too Dumb to Die", Kim McKinzey highlights some issues associated with the fact that, crudely over-stated, the laws relating to the death penalty for murder in some States of the US Union allow that the actions of mentally retarded murderers can be excused because they are likely to have failed to understand the implications of their actions.

So how to determine whether someone is mentally retarded or not? As McKinzey shows, a host of tests, ranging from the Wechsler Intelligence Tests to the Vineland Social Maturity scales, have been deployed by forensic Psychologists, their relevance disputed in court, and the results compared with "common sense" assessments of "real life" behaviour.

But behind such activities lies another set of disputes: How good are the samples on which the norms are based? If prospective participants in such studies have exercised their "informed consent"-based rights not to participate, what effect has that had on the norms? What is the effect of the *date* on which the norms were collected? (Judged against yesterday's norms one should die; yet, given today's norms, one may live). And what statistical procedures have been deployed to compile the norms? (As Dockrell[1] has shown, the IQ of the same person on the same test judged

against the same norming sample can vary dramatically depending purely on the assumptions made by the *statistician* who processed the data.)

Yet even such questions seem somehow to miss the point. That point has two facets – one to do with ethics; the other with competence.

One of the most surprising conclusions to emerge from both our own work and that of others (such as Donald Schon), is that incompetence in modern society stems above all from an inability and unwillingness to engage with the wider social forces which *primarily* determine behaviour and thus what people *can* do in their jobs. In this case, this implies that we, as psychologists, need to get together with others (perhaps through our professional organisations) to influence the social and legal contexts in which we work instead of accepting those contexts as givens. Furthermore, unless we do this we *cannot* behave ethically.

In the second chapter we return to Jim Flynn for a remarkable discussion of issues typically touched on only superficially in discussions of topics having to do with such things as "fairness in testing" and, in particular, the assumed viability and ethics of a meritocracy (which often informs discussions of "fairness in testing").

1. Dockrell, W. B. (1989). Extreme scores on the WISC-R. *Bulletin of the International Test Commission, 28*, April, 1-7.

Chapter 24

Too Dumb to Die: Mental Retardation Meets the Death Penalty*

R. Kim McKinzey

Abstract

In Atkins v. Virginia (2002), the US Supreme Court held that executing the mentally retarded is unconstitutional. In a capital, death penalty case, a hearing must therefore be held sometime before sentencing or trial to determine whether or not the defendant is mentally retarded. An Atkins case study is presented, wherein the issues involved are discussed. These issues include: timing of the hearing, burden of proof, malingering, data gathering, and measurements of intelligence and adaptivity.

The author has also prepared an update summarizing subsequent developments. This can be found at http://wpe.info/papers_table.html

******

On the 16th of August 1996, Daryl Renard Atkins kidnapped, robbed, and shot Eric Nesbitt. Atkins was convicted of a capital crime, and the case went to the mitigation (penalty) phase[24.1], where the jury considered a defense claim of mental retardation.

According to the eventual US Supreme Court decision ("Atkins v Va.," 2002)[24.2], Evan Nelson, Ph.D.,[24.3] a defense psychologist, testified that Atkins has mild mental retardation, citing a Wechsler Adult Intelligence Scale, third edition (WAIS-III[24.4]) Full Scale Intelligence Quotient (FSIQ) of 59.[24.5] While an adaptation measure was not used, a review of Atkins' history showed a "lack of success in pretty much every domain of his life" (Judge Scalia's dissent, p. 2, quoting Dr. Nelson's testimony).

* Like many other chapters in this book, this article has, for some time, been available at http://wpe.info/papers_table.html

After an appeal, a second mitigation/sentencing jury was assembled, and Dr. Nelson repeated his testimony. This time, Stanton Samenow, Ph.D.[24.6], a prosecution psychologist, disputed the claim, arguing (without the use of an IQ test) that Atkins was of average intelligence. Atkins' poor academic performance was due to his choosing not to attend, an early symptom of his Antisocial Personality Disorder (APD[24.7]). The second jury also sentenced Atkins to death, perhaps because they also heard of his 16 prior convictions for robbery, attempted robbery, abduction, gun use, and maiming, including graphic descriptions from former victims, one of whom was pistol-whipped and shot.[24.8]

In the required appeals, the appellate judges preferred Dr. Nelson's opinion. Citing the then predominating US Supreme Court's decision on the topic ("Penry v Lynaugh," 1989) the majority of the Virginia Supreme Court[24.9] thought Atkins acceptable for execution. Two dissenting judges (Justices Hassell and Koontz) thought Dr. Samenow's opinion was "incredulous as a matter of law" and argued that Atkins should be spared execution solely due to his mental retardation. On 20th June 2002, the US Supreme Court reversed itself, and decided that the time had come to end the execution of people with Mental Retardation (MR), finding it an "excessive," "cruel and unusual punishment" and in violation of the Constitution's Eight Amendment[24.10].

Clinical definitions of mental retardation require not only subaverage intellectual functioning[24.11], but also significant limitations in adaptive skills such as communication, self-care, and self-direction that become manifest before age 18. Mentally retarded persons frequently know the difference between right and wrong and are competent to stand trial. Because of their impairments, however, by definition they have diminished capacities to understand and process information, to communicate, to abstract from mistakes and learn from experience, to engage in logical reasoning, to control impulses, and to understand the reactions of others. There is no evidence that they are more likely to engage in criminal conduct than others, but there is abundant evidence that they often act on impulse rather than pursuant to a premeditated plan, and that in group settings they are followers rather than leaders. Their deficiencies do not warrant an exemption for criminal sanctions, but they do diminish their personal culpability. (page 13)

The opinion offered no other guidelines on the many resultant issues: "We leave to the State(s) the task of developing appropriate ways to enforce the constitutional restrictions upon its execution of sentences." (page 12)

Naturally, the condemned across the country began appealing their sentences. If a person on death row has MR, the sentence must be converted – but, to what? Life Without Parole (LWOP)? Life? Something else? And who qualifies for this life-saving diagnosis? Just those with a pre-crime diagnosis? How about those who developed MR after the crime?

As Atkins made clear, the judicial finding that a person has MR is obviously quite different from the diagnosis of MR. Theoretically, mental health professionals make diagnoses based on the combination of the scientific literature and the condition of the patient, while triers-of-fact (TOF) must consider only the testimony of the diagnosing professionals. If the latter disagree, the conflict must be resolved by the TOF. For that to happen, testimony must be heard in a hearing.

Who is the TOF? A judge? Federal or State? Appellate or trial? A jury convened to decide? Are they to be Death-Qualified? Who can be witnesses? Anyone, or just professionals qualified to diagnose MR? Who are those professionals, and what qualifies them to diagnose a life-saving condition?

What's the burden of proof, and who has it? If the court, prosecution, and defense can all pay for experts, can all three examine the person? Can a defendant/appellant in this situation be ordered to cooperate? What confidentiality rules apply? Can the results of the examination be used in other hearings?

The Atkins ruling came in the midst of already-ongoing cases, producing more issues. When is the hearing to be held? Pre-trial or mitigation? If a person with MR cannot be executed, can the person still be charged with a capital crime? Does that change the funding of the trial or the requirements for the qualifications of the trial attorneys or the jury's Death Qualification? In the absence of clear legislative and judicial guidelines, can a judge ruling on an issue be sure of not being reversed?[24.12]

The issues confronting the testifying professionals are equally numerous. Do the professionals need any special qualifications? Should they be held to a higher standard in a life-or-death diagnosis? Is there any esoteric literature on the topic? What procedures/tests/measurements should be used? Should those procedures/tests/measurements be held to a higher standard? What if the patient meets one MR standard and not another? And what about the issue of malingering: "One need only to read the definitions of mental retardation…to realize that the symptoms of this condition can readily be feigned." (Justice Scalia's dissent, p. 17).

Into this morass of questions came a capital defendant, Jose Lopez, whose case is presented as an example of how the participants in one California county court trial handled the issues. In the words of Margaret Talbot (Talbot, 29 June 2003), was Lopez "too dumb to die"?

Jose Lopez, A Case Study

Facts of the case. According to the police reports, Jose Lopez[24.13] gathered with five other young (one a minor) men in a rural California town on a summer's day in 2001. They had been feuding with a rival gang, and one of their members had recently been stabbed. Determined to gain revenge, the six men got into a recently stolen car and drove to the assailant's house. Spotting him and two other men on the porch, one of the men in the car opened fire, killing the assailant and wounding one of the other two. The six men drove away, hijacking a truck when the car malfunctioned.

The six men were rapidly identified, and Lopez confessed to being in the car. Some of the other men also confessed, and named Lopez as the driver–although he has no driver's license, he was the best driver amongst them. All agreed he was not the shooter. In his explicit, videotaped confession[24.14], he noted the shooting was pre-meditated: "When we all get together like that, we don't just get together to talk, ya know?" He described the getaway[24.15] route in detail: To orient the detectives, he drew a map, and mentioned as landmarks seven stores, three streets, seven changes of directions, and two stop signs. He calculated the number of bullets used, explaining that 6+6+2=14. He described the three vehicles and two guns involved. He explained how he had driven the truck he had just hijacked to find the now 16-year old mother of his one-year old son,[24.16] with whom he had been living a month prior, and had an emotional fight with her. He gave details of the men's family relations to each other. He became visibly emotional when he realized he had both just confessed and fingered his fellow gang members, and asked for protection from their wrath. There were no obvious problems with his face (dysmorphia), articulation (dysarthria), or vocabulary.

The prosecution decided to bring capital charges against the five adult men, with the remaining minor cooperating with the prosecution. The case was assigned to a county judge and defense attorneys qualified for a capital case.[24.17]

In doing the necessary background investigation[24.18], the defense discovered that the 19 year-old Lopez had had behavioral difficulties for most of his life. More to the point, he had been in Special Education for years, and had his IQ tested twice (see Table 24.1 for a listing of measures and results). Since one of the IQ evaluations suggested Lopez has MR, the defense sought a hearing under Atkins.

Childhood evaluations. Coming from a Mexican immigrant family, Lopez entered school speaking little English. He was kept in Kindergarten an extra year, where he was behaviorally aggressive. At the end of the year, in 1990, he was examined by a Spanish-speaking School Psychologist[24.19], Mr. A, who gave the Wechsler Intelligence Scale for Children, Revised version, Mexican edition (WISC-R-M), Peabody Picture Vocabulary Test, Revised version (PPVT-R), Spanish Woodcock-Johnson, Dos Amigos Verbal Language Scales, and the Spanish version of the PPVT-R (PPVT-R-S). Not surprisingly, Lopez did very badly on all of the English-based tests and much better on all of the Spanish-language tests. His WISC-R-M FSIQ was 91, in the Average range of intellectual functioning. His equivalent PPVT-R-S IQ score was 82, in the Low Average range. Since his academic functioning was well below that of his IQ, he qualified for more intense teaching resources, in the form of Special Education.

Placed in Special Education for first grade, the young Lopez was given help in increasing his English fluency. He passed the grade, and he was taken out of Special Education. Although an indifferent student, he passed each succeeding year. but fell behind again, and was assigned a tutor, in the person of Mr. B, another Spanish-speaking School Psychologist.

In late 1995, a teacher wrote that Lopez "is having difficulty adjusting to an English-only classroom. He becomes easily frustrated and acts out[24.20]… He has spent the majority of his time in in-house suspension." Mr. B wrote that Lopez had "significant behavioral, family, and emotional problems. In terms of his behavioral difficulties, [he] demonstrates oppositional defiance, argumentative and antisocial tendencies." At 12, he had already been expelled for carrying a knife.

The next year did not go better, and Lopez was placed in independent study. At the end of the year (1996), Mr. B again evaluated Lopez's IQ. The WISC-III FSIQ was 55, in the Mildly Mentally Retarded range. However, Mr. B wrote, Lopez was clearly not trying his best, being suspicious, defiant, and impulsive. Mr. B characterized the test results as being "somewhat depressed."

Psychologists have known for decades (Heaton, Smith, Lehman, & Vogt, 1978; Ziskin, 1995) that patients do not always try their best on IQ tests, especially in cases where some benefit or legality is involved. Research efforts to find an accurate method of detecting such malingered performances resulted in many failures. Psychologists took four approaches:

1. They tried to use personality tests that had already established validity measures. This method was rapidly disproven (Heaton et al., 1978).
2. They tried to use their own notions of what constituted a consistent pattern of scores, arguing that they could "just tell" when a given set of scores was faked. This method was eventually disproven (Faust, Hart, & Guilmette, 1988; Faust, Hart, Guilmette, & Arkes, 1988).[24.21]
3. They tried to use stand-alone tests devised especially for detecting neuropsychological faking. Although two of these (McKinzey, 5 April 2003; Tombaugh, 1996) now have acceptable levels of accuracy, their use with patients with MR can only be considered promising (see the Author's Update[24.22] for details on this fast moving area of research). As yet, no one has *clearly* demonstrated that a person attempting to get a deliberately faked IQ score will also fake these stand-alone measures.
4. They tried to devise formulas using the test's own within-test scores to identify faked scores. So far, only two tests have had such within-test formulas cross-validated, the Wechsler Adult Intelligence Scale, Revised & Third versions (WAIS-R & WAIS-III) and the Raven Standard Progressive Matrices. The two WAIS formulas (Mittenberg et al., 2001; Mittenberg, Theroux-Fichera, Zielinski, & Heilbronner, 1995), although promising, have not been fully cross-validated (see the Author's Update for citations and reviews of this rapidly moving area of research). The Raven formula has been cross-validated, and found to have a false negative rate of 29% and false positive rate[24.23] of 2.5-10% (McKinzey, Podd, Krehbiel, & Raven, 1999).

In the first of several counter-intuitive decisions, Mr. B did not increase his rapport with Lopez to obtain a valid IQ score. Instead, he simply observed that Lopez's academic performance was now in line with his tested IQ, and he no longer qualified for any special assistance. Then, Mr. B did not refer Lopez to the local Regional Center, as he was required to do if he thought MR was a reasonable diagnosis.

The next two years did not show any improvement, and Lopez was in Juvenile Hall by 1998. He got a 14-year old teenager pregnant in 1999 and lived with her briefly in 2001. He was arrested soon afterward for the shooting.

The judicial decisions. Faced with few definitive legislative or judicial guidelines, the trial judge decided the Atkins hearing would be pre-trial, himself as TOF, with burden of proof at preponderance (more likely than not). The prosecution was allowed to examine the defendant, which was, counter-intuitively, declined,[24.24] making a variety of issues moot. The court appointed two psychologists to examine the defendant and make recommendations.

The forensic evaluations. Dr. C was retained by the defense. Deeming Lopez's English-language capacity to be adequate, the doctor gave him two IQ tests requiring English fluency, the WAIS-III, which yielded an FSIQ of 67, and the Shipley Institute of Living Scale, which yielded an IQ of 52.

As a check, the doctor also gave the Raven Standard Progressive Matrices, a relatively culture-free IQ test that requires no English ability. The test yielded a standard score of below 70 (to get a more accurate estimate of <70 IQ, another version of the test, the Coloured Progressive Matrices, must be used, which the doctor did not), even when scored correctly (which the doctor did not). For some reason, the doctor did not attend to the manual's instructions (Raven, Raven, & Court, 2000), which require the test-taker to get all of the first three extremely simple items correct before continuing with the test.

More importantly, the doctor also did not attend to the manual's citation of the malingering index for the Raven[24.25]. When scored, this index demonstrated Lopez to be faking his low IQ. Even more importantly, none of the other doctors except the one called by the prosecution noticed this finding either. Dr. C gave no test of adaptation, but noted Lopez's educational problems and deficits in social, interpersonal, and home-living skills. He offered a diagnosis of Mild MR.

Dr. D was appointed by the court. He gave the WAIS-III to Lopez two months after Dr. C gave the same test, and obtained a standard score of 65. He obtained a history of Lopez playing organized baseball at age 12.

Dr. D also gave a test of adaptation, the Vineland Adaptive Behavior Scales (Sparrow, Balla, & Cicchetti, 1984). Such tests use a standardized set of questions and norms to survey the patient's functioning in a variety of areas, such as academics, self-care, communication, safety,

socializing, and misbehavior. The norms compare the patient's scores to those obtained by people with and without MR. Lopez's Vineland scores were all quite low, similar to those obtained by people with MR. Dr. D therefore offered a diagnosis of MR.

Dr. E was appointed by the court. He gave no IQ or adaptive tests of his own, relying instead on those reported by Drs. C and D. He agreed with them in their diagnoses of MR.

The hearing. Drs. C, D, and E were called by the defense, each reporting their conclusions that Lopez was MR. None had heard of the Raven validity formula, and simply pled ignorance when confronted. None changed his opinion when told of the first evaluation and outcome of the Raven formula.

Mr. B was also called by the defense. He now agreed Lopez was MR, pointing to the new IQ testing as confirming his opinion. He further opined that the WISC-R-M produced inflated scores, citing his own experience with the test.

Mr. A was never located for testimony.

Dr. F was called as rebuttal by the prosecution. He opined that Lopez did not meet any definition of MR, pointing to the first evaluation and its two above-70 IQ scores. He noted the faked Raven, the complete lack of previous diagnosis (despite years of psychological attention), and Lopez's normal level of adaptation. He defended the WISC-R-M and PPVT-M, pointing out that no scientific literature existed demonstrating them to yield inflated scores. He also pointed out that the Vineland's norms stop at age 18, and Lopez was 19 when tested. The Vineland is largely self-report, and has no method of detecting misleading, malingered answers.

However, there's a more important problem with the Vineland (and all the other adaptivity scales): They are not properly validated for the task being asked of them.

The Atkins decision requires experts and TOFs to distinguish between people with and without MR who have been accused of a violent crime. Like both Atkins and Lopez, many of these people will have childhood histories of violence and/or willfulness. The current adaptive tests have not been designed to distinguish between willfulness and disability.

Some examples: One Vineland item asks if the patient cleans his room. The item does not distinguish between the patient who does not clean his room because he is too intellectually limited and one who willfully, parasitically, in Lopez' words, "lets my bitch do it." Although one item asks if the patient is dating (which Dr. D decided Lopez did not

do), there is no item for "is sexually active," "is a member of a gang," "drives a car", "can shoot a gun", or "has planned a crime with others".

The authors of one adaptive test, the Adaptive Behavior Assessment System (ABAS) (Harrison & Oakland, 2000) asked a simple question: How accurate was the ABAS? They compared scores of people (5-18 years old) with Mild MR and those without MR, with and without serious childhood misbehavior. Using a variety of comparisons and decision rules, they found 32% of their *normal* sample could be mislabled as MR, while 50% of the people with Mild MR got normal scores (see Table 5.31, page 89). Adults (ages 17-72) without MR were misclassified 17% of the time (see. Table 5.31, page 90). Children (ages 6-21) without MR but with behavior disorders were misclassified 73% of the time (see Table 5.33, page 93). Those with emotional disturbances (ages 5-18) were misclassified 70% of the time. There were no samples of the parasitic, willfully self-indulgent impulsive lifestyles typical of people with Antisocial Personality Disorder.

No one has even asked if the Vineland has similar misclassification rates.

The ruling. The judge decided the defendant had met his burden, and the prosecution was prevented from pursuing the death penalty.

Discussion

Atkins hearings will take predictable courses. The defense experts will find ways of explaining away normal IQs and adaptive functioning and produce scores in the MR range. The prosecution experts will deride the new test scores and argue the defendant is merely a malingering crook. After a battle of experts, the judge will utilize a terrible calculus: If the judge rules the defendant has MR, the matter ends without appeal, and much money is saved by avoiding a death penalty trial. If the judge rules the defendant does not have MR, the ruling automatically becomes appealable and a long series of state and federal judges will feel free to override the finding. Using juries to make the decision will produce more MR decisions but raise far more questions about selection. Will the jury be Death Qualified? Can it include people with MR ? People with MR in the family? People with professional experience working with MR? Each trial will require a long series of appellate cases to resolve.

It will take another decade for psychologists to improve their tools enough to be adequate. In an Atkins decision, a one point difference

between two scores is life or death! The issues of accuracy rates and malingering formulas will dog tests of both IQ and adaptivity.

As yet unexplored is the issue of interview source. Adaptivity test scores differ with who is answering the questions (and probably who is asking them). What will a defendant's mother say when the prosecution's expert asks if the defendant cleaned his room? Will she remember what she told the defense's expert?

What a situation for a mother!

Table 24.1. *Lopez' Test Results*

Administrator	Date	Test	Score	Range
Mr. A	5/90	WISC-R-M FSIQ	91	Average
		PPVT-R Spanish	82	Low Average
Mr. B	6/96	WISC-III FSIQ	55	Mild MR
Dr. C	9/02	WAIS-III FSIQ	67	Mild MR
		Raven	<70	Mild MR
		Shipley ILS	52	Mild MR
Dr. D	11/02	WAIS-III	65	Mild MR

Note. Date of birth: 1/83. Date of shooting: 7/01. Scores are Standard, mean = 100, SD = 15. Mr. A & B are Spanish-speaking, MA-level school psychologists. Drs. C & D are Ph.D.-level clinical psychologists. WISC-R-M is the Wechsler Intelligence Scale for Children, Revised, Mexican edition. WISC-III is the WISC, third edition. WAIS-III is the Wechsler Adult Intelligence Scale, third edition. FSIQ is the Full Scale Intelligence Quotient. PPVT-R Spanish is the Peabody Picture Vocabulary Test, Revised, Spanish version. Raven is the Raven Standard Progressive Matrices. Shipley ILS is the Shipley Institute of Living Scale.

Notes

24.1 In those US states with the death penalty, the defendant must first be convicted of a killing accompanied by an additional felony (Special Circumstances, or Aggravating Factors), such as a related killing or robbery. The case then enters the penalty (or mitigation) phase, wherein the same jury (which are specially screened for their willingness to recommend death, so called Death Qualification) then hears of *any* factors that might lessen (mitigate) the horrendousness of the crime. If any such factors are found, the jury may recommend Life Without Parole (LWOP) instead of death. Examples of such factors are mental retardation, dementia, youth, lack of previous record, childhood abuse, relative lack of culpability

24.2 The opinion can be downloaded at: http://www.supremecourtus.gov/opinions/01slipopinion.html

24.3 According to his website (http://www.psylaw.com/dr_enelson.htm), Dr. Nelson obtained his doctorate from the Univ. of North Carolina in 1991. He worked for the forensic unit of a state hospital for three years before going into private practice, specializing in legal referrals. He has published on informed consent in insanity trials.

24.4 A note on test names: Tests (and sometimes content) must be updated every few years. The second generation of the test is then designated as 2, II, or Revised, with the third generation designated as III. This can sometimes be misleading: in the case of the WAIS, the third edition is actually the fourth version, having been preceded by the Wechsler Bellevue.

24.5 Dr. Nelson concluded that this score was not an "aberration, malingered result, or invalid," since Atkins' limited intellect had been consistently present his entire life. See Atkins v Virginia, footnote 5.

24.6 According to his website (http://www.samenow.com), Dr. Samenow got his doctorate from the Univ. Michigan in 1968. For eight years, he worked with Samuel Yochelson, producing an important text on *The Criminal Personality*. He went into private practice in 1978, focusing on legal referrals. He published *Inside the Criminal Mind* in 1984, *Before It's Too Late in 1989, Straight Talk About Criminals* in 1998, and *In the Best Interest of the Child* in 2002.

24.7 The definition of APD can be found in DSM-IV-TR (American Psychiatric Association, 2000), or http://www.behavenet.com/capsules/disorders/antisocialpd.htm

24.8 While relevant, the issues of death-qualification of juries and victim impact statements are well beyond the scope of this paper.

24.9 The Virginia opinions can be found at: http://www.courts.state.va.us/scndex.htm

24.10 The US Constitution's full text can be found at: http://www.archives.gov/exhibit_hall/charters_of_freedom/constitution/constitution.html

24.11 All of the several MR definitions require measured IQ to be about two standard deviations below the measure's mean. For most tests, the cutoff is about 70, with mean 100 and Average IQ being 90-110.

24.12 The American Association on Mental Retardation (AAMR) has a website offering a variety of opinion papers on some of these issues: http://www.aamr.org

24.13 Although the case materials are now public record, the name has been changed for privacy. Jose Lopez is meant to be the Hispanic version of John Doe.

24.14 Lopez waived his Miranda rights. No hearing was held to determine his competency to do so. While relevant, the literature on competency to waive Miranda is beyond the scope of this paper. The literature on false confession is not relevant to *this* case study.

24.15 When the gang finished their getaway, the only member with a car balked at taking the rest home, since they wouldn't contribute gas money.

24.16 The mother of this then-14 year old pregnant girl was not charged with failure to protect.

24.17 For the CA guidelines on such attorney qualifications: http://www.courtinfo.ca.gov/rules/titlefour/title4-13.htm

24.18 For guidelines on doing such a background investigation, see: http://www.criminaljustice.org/public.nsf/941a6d5b3ad55cd485256b05008143fd/bee3ff4450880bb485256704006793eb?OpenDocument

24.19 CA School Psychologists are Master's level psychologists not allowed to practice independently. For details of their duties, see: http://www-gse.berkeley.edu/program/sp/html/what_is_a_school_psych_.html

24.20 "Acting out" is a euphemism for rule-breaking misbehavior.

24.21 I foresee this argument becoming the method of choice for psychologists unwilling to use validated methods. It will appear when the defendant's IQ scores are close to each other on multiple testing. The psychologist will argue, sans scientific proof, that malingerers simply cannot manage to get such consistent results.

24.22 http://wpe.info/vault/td2/tdtdau.pdf

24.23 The false negative rate refers to the percentage of the time the condition being tested for (e.g., malingering), is missed. The false positive rate refers to the percentage of the time the condition being tested for is falsely detected in people without the condition. For more test accuracy definitions, see: http://wpe.info/2x2table.pdf

24.24 The reasons the prosecution might decline to examine the defendant are beyond the scope of this article.

24.25 The discerning reader will notice that I have not reported Lopez' scores on the WAIS-III index. When this index is, in my opinion, adequately cross-validated, I will add the outcome in the Author's Update.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders, Fourth Edition, Text Revision*. Washington, DC: Author.

Atkins v. Virginia 536 U.S. 304 (2002).

Faust, D., Hart, K. J., & Guilmette, T. J. (1988). Pediatric malingering: The capacity of children to fake believable deficits on neuropsychological testing. *Journal of Consulting & Clinical Psychology, 56*(4), 578-582.

Faust, D., Hart, K. J., Guilmette, T. J., & Arkes, H. R. (1988). Neuropsychologists' capacity to detect adolescent malingerers. *Professional Psychology: Research & Practice, 19*(5), 508-515.

Harrison, P., & Oakland, T. (2000). *Adaptive Behavior Assessment System: Manual*. San Antonio: Psychological Corporation.

Heaton, R. K., Smith, H. H., Lehman, R. A., & Vogt, A. T. (1978). Prospects for faking believable deficits on neuropsychological testing. *Journal of Consulting & Clinical Psychology, 46*(5), 892-900.

McKinzey, R. K. (4/5/03). The current accuracy rates of the Word Memory Test. *WebPsychEmpiricist*. Retrieved April 5, 2003, from http://wpe.info/papers_table.html

McKinzey, R. K., Podd, M. H., Krehbiel, M. A., & Raven, J. (1999). Detection of malingering on the Raven Progressive Matrices: A cross-validation. *British Journal of Clinical Psychology, 38*(3), 435-439.

Mittenberg, W., Theroux, S., Aguila-Puentes, G., Bianchini, K., Greve, K., & Rayls, K. R. (2001). Identification of malingered head injury on the Wechsler Adult Intelligence Scale–3rd edition. *The Clinical Neuropsychologist, 15*(4), 440-445.

Mittenberg, W., Theroux-Fichera, S., Zielinski, R., & Heilbronner, R. L. (1995). Identification of malingered head injury on the Wechsler Adult Intelligence Scale–Revised. *Professional Psychology: Research & Practice, 26*(5), 491-498.

Penry v. Lynaugh 492 U.S. 302 (1989).

Raven, J., Raven, J. C., & Court, J. H. (2000). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3: Standard Progressive Matrices*. San Antonio, TX: Harcourt Assessment.

Sparrow, S., Balla, D., & Cicchetti, D. (1984). *Vineland Adaptive Behavior Scales: Interview Edition Survey Form Manual*. Circle Pines: MI.

Talbot, M. (6/29/03). The executioner's IQ test. *The New York Times Magazine*. Retrieved June 29, 2003, from http://www.newamerica.net/index.cfm?pg=article&pubID=1275

Tombaugh, T. N. (1996). *Test of Memory Malingering: TOMM*. Niagara Falls, NY: MHS.

Ziskin, J. (1995). *Coping with psychiatric and psychological testimony* (5 ed.). Los Angeles, CA: Law and Psychology Press.

Chapter 25

Excerpts from: *"How to Defend Humane Ideals"**

Jim Flynn

Abstract

This paper presents two sets of extracts from the author's book *How to Defend Humane Ideals.* The first set fundamentally challenges a great deal of "politically correct" thinking on bias in testing and social policy. One of the main organising constructs is "ethnicity as an information bearing characteristic". It is argued that it is naïve to think that individualistic assessment can meaningfully replace the use of the information bearing capacity of group differences in public policy. The second set of extracts clearly demonstrate that those who imagine that a single-factor meritocracy can be combined with gross differences in material standard of living are simply not thinking clearly. Resolution of the dilemma posed by the tension between meritocracy and egalitarianism behoves psychologists to find ways of identifying, developing, and rewarding multiple talents on the one hand and ways of understanding and intervening in the sociocybernetic forces which prevent people enacting their values (both individually and collectively) on the other.

******

* A version of this chapter has for some time been available in electronic form on the *Web Psych Empiricis*t http://www.wpe.info/papers_table.html

Editorial Introduction

By entitling an article in The American Psychologist "*Searching for justice: The discovery of IQ gains over time*", Flynn (1999) makes it clear that his research into the, at the time, almost entirely unsuspected effects of environment on IQ was driven by ethical considerations. In fact, defending human rights and humane ideals had, by this time, been a major theme in Flynn's life. But what he does not say in his article is that, at the very time it was published, he was engaged in working up a major book entitled "*How to Defend Humane Ideals*". Why should a scientist allow his work to be driven by humane ideals instead of by questions derived from previous research?

"*How to Defend Humane Ideals*" is, in reality, a tour de force drawing together what philosophers from Plato onwards have had to say on the topic, explaining why it is important to defend humane ideals against their anti-humane opponents, and setting out how to do it.

But, given Flynn's commitment to humane ideals, and especially in the context of much contemporary group think, it is striking to find in the book a trenchant critique of much widely accepted and "politically correct" "thinking" on such topics as bias, prejudice, stereotyping, and the concept of meritocracy which informs many discussions of "fairness" in testing.

I asked Flynn to make a selection from his book (and update it) that would illustrate his arguments relating to bias and meritocracy. What follows is, in fact, of particular importance to us here because, although Flynn does not say so (but which does not mean that he would not have done so if he had been asked), behind his observations on these topics lies a network of questions which psychologists have a moral responsibility to address.

*****

Excerpt 1: Race as an information-bearing trait that disadvantages blacks - with some more recent data appended.

The humane-egalitarian ideal of social justice presented herein rests on sympathy for people in general, operationalized by leveling differences that are the effects of fortune. It includes affirmative action as a compensation for the luck of group membership, the welfare state as a compensation for the luck of genes, and redistribution of wealth

as a compensation for the luck of personal circumstance. First, I will defend affirmative action for blacks living in contemporary America; then I will defend the ideal of equalizing environments, toward which the welfare state and redistribution of wealth are steps, against Herrnstein and Murray. These critics happen to be American, but the substance of their case, the meritocracy thesis, has been put forward by opponents of equality throughout the European world ever since the dawn of the industrial revolution.

Blacks as a Disadvantaged Group

Social science collects evidence on group differences. Sometimes it shows that putative differences between black and white Americans are illusions based on ignorance or bias. That can advantage blacks. Sometimes it shows that differences are real and must be accepted by all rational agents. As we shall see, if those agents are truly rational, they will then make certain choices to the disadvantage of blacks. Social science can do nothing about this except conceal the truth, and that it must not do. However, its practitioners must not close their eyes to the consequences of their science. They often say, "It makes no difference if we show that blacks on average are genetically inferior for intelligence, are less prudent and self-disciplined than whites, tend to be more criminal. Only a biased person will discriminate against people according to their group membership rather than judging them by their individual traits." I will show that this last assertion is false.

Social science also attempts to measure how much bias exists. Herrnstein and Murray (1994, 506) believe that while undeniably some bigotry still exists, the vast majority of Americans are fair-minded and free of racial prejudice. Rather than challenging that conclusion, I will treat it as a window of opportunity. If we can show that even in the absence of bias, individual blacks are gravely disadvantaged simply because of their group membership, that might be the strongest possible case for affirmative action. Therefore, the organizing concept of this analysis will not be racial bias but the cost of information.

Levin (1991) points out that race can be an information-bearing trait. He cites a variety of sources as showing that one black male in four is incarcerated at some time for the commission of a felony, while the rate for white males is only about 3 per cent, and that a black male is ten times more likely than his white counterpart to be a criminal (Berger, 1987; Hindelang, 1978; Rushton, 1988; V.S. News and World Report,

1988; Wilson & Herrnstein, 1985). He endorses the practice of the New Jersey police of stopping young black males in expensive new cars for random drug searches. After all, police resources are stretched, and their efficiency in controlling the drug traffic is maximized by information that enhances the probability of finding illegal drugs. The dividends of targeting blacks extend to other areas of crime prevention. As police officer Mark Furhman of O. J. Simpson fame put it, if a black man is driving a Porsche and wearing a suit that costs less than $100, you stop him on the assumption that the car may be stolen. Anyone who listens to a police radio will discover that blacks who walk through a white neighborhood are labeled suspicious, while whites in a black neighborhood go without remark.

It is rational for police to use race as a low-cost information bearer to enhance their efficiency. Is it rational for blacks to resent this and take steps to make the information more expensive? A few examples may help. Irish Americans have a rate of alcoholism well above that of most ethnic groups. When resources are stretched, as always, and the highway patrol is conducting random checks for drunken drivers, they would do well to stop only Irish male drivers, particularly where Irish are heavily concentrated. The problem is that they cannot be identified by appearance, and stopping all drivers to verify whether or not they were Irish would be self-defeating. Irish could be forced, and everyone else forbidden, to drive green cars, but that law might be evaded. The rational solution would be shamrocks indelibly tattooed on the foreheads of all Irish males, perhaps luminescent at night. There would be a cost in this, but it could be shifted to the Irish themselves. Levin also notes that people associate insider trading with Jewish-Americans. This association may not be based on evidence, and the resources of the Securities and Exchange Commission may not be stretched. But if those conditions hold, the utility of Stars of David becomes obvious.

Every black knows that Irish and Jewish Americans would raise the cost of collecting this sort of information to a prohibitive level by political action of the most impassioned sort. Their own efforts have had mainly a cosmetic effect: police omit race from the formula of criminal profiles but continue to use it in practice. Therefore, added to whatever humiliation blacks feel at random searches, there is a sense of overwhelming political impotence. Since blacks cannot use politics to raise the cost, it is rational to pursue other means both individually and collectively?

On the individual level, those stopped for random searches will tend toward noncooperation, verbal abuse, attempts at escape with

attendant low-level violence. The police, being rational agents, are likely to anticipate this and resort to preventive measures, that is, they are more likely to handle and search black suspects roughly, even to perpetrate the occasional beating, hoping to intimidate and achieve control. The black community can collectively increase costs to the police by making it clear that if black suspects are abused, there is an ever-present chance of riot. You now have a significant level of random violence between police and black males, but there need be no animosity or real bias on either side. Black males may not dislike police simply because they are police nor police blacks simply because they are black. Both sides may recognize that the other's behavior is simply a rational response to objective group differences. Stove (1995, 95) adds a point that takes us from theory back to reality, namely, that even rational behavior, just so long as it inflicts injury, can engender strong negative feelings between groups. It can indeed.

If negative racial profiles of blacks are rational, we would expect them to be used by blacks as well as whites and to be used extensively. Both white and black landlords are more reluctant to rent to young black males - after all, 25 percent of them are convicted felons and who is to know which. Both white and black banks are more likely to lend money to entrepreneurs outside the ghetto - those within are seen as greater risks. Both white and black car dealers ask and get higher prices from blacks than whites - they see them as worse informed and less confident about bargaining. When shown photographs of blacks distinguished only by lighter or darker skin color, both whites and blacks identified blackness with poverty, aggressiveness, lack of intelligence, lack of education, and unattractiveness. Blacker males were also seen as criminal and ostentatious (Ayers & Siegelman, 1995; Maddox & Gray, 2002).

Two pieces of data are particularly shattering. Since 1941, uniformed police have shot 23 black policemen working undercover in New York City alone; no white policeman working undercover has ever been shot by a colleague. When Bertrand and Mullainathan (2003) sent 5,000 resumés randomly assigned to either white or black sounding names (Emily and Greg or Lakisha and Jamal) to 1,250 employers who had placed help-wanted ads, the white names received 50 percent more call-backs. Average white applicants got many more call-backs than highly skilled black applicants, indeed, black applicants were treated as if their qualifications did not matter: high quality resumés got no more calls than average resumés. Human resources managers consulted beforehand were

stunned. They believed that the results would reflect employers hungry for qualified minority applicants and aggressively seeking diversity.

Note that the application of these profiles does much to explain the reluctance of white males to marry black women. After all, their children would be socially classified as black. Why would a white man want to saddle his children with that when he has so much choice? For a white male to want her as a spouse, a black woman must have an appeal well beyond that of an Hispanic or Asian woman.

Excerpt II: On meritocracy.

Humane-egalitarian ideals may include a coherent concept of justice, but can they accommodate what human genetics and social dynamics tell us about certain group differences? Herrnstein and Murray claim that they cannot and use the meritocracy thesis as the vehicle for their argument. I will rebut the meritocracy thesis and use that rebuttal to extract a bonus: a deeper insight into the dynamics of humane-egalitarian ideals. Herrnstein and Murray (1994, l05, 109, & 510) state the meritocracy thesis in four propositions: (l) if differences in mental abilities are inherited, and (2) if success requires those abilities, and (3) if earnings and prestige depend on success, (4) then social standing (which reflects earnings and prestige) will be based to some extent on inherited differences between people. They imagine a United States that has magically made good on "the contemporary ideal of equality". First, every child has equal environmental quality insofar as environment affects intelligence. Second, each person can go as far as talent and hard work can take him or her with neither social background, nor ethnicity, nor lack of money barring the way.

Herrnstein and Murray (1994, 91, 105-115, & 509-520) believe that America has realized the humane-egalitarian ideal in practice to a significant degree. The irony is that, insofar as it is rea1ized, America approaches a kind of caste society egalitarians would loathe. If environmental inequality is diminished, intelligence differences between individuals increasingly reflect genetic differences. If privilege is diminished, intelligence or IQ becomes an enhanced factor in social mobility, so that upper-class occupations become filled by the bright and lower-class occupations by the not bright. Genes for intelligence become more and more segregated by class. There is an elite class with good genes for IQ whose children tend to replicate their parents' high status because of luck in life's lottery, that is, because they inherit their parents' good genes. There is a large underclass with bad genes for IQ whose children suffer from cognitive disadvantage at birth and find it difficult to escape low status.

The meritocracy thesis strikes at the very heart of the human-egalitarian ideal. That ideal is revealed to be counter-productive in practice. The abolition of inequality and privilege produces a class-equals-caste society with high status the inheritance of a few, dependency and low status the inheritance of many. How little this vision will appeal will vary from person to person, but it is safe to say that countless idealistic men and women did not lay down their lives for this.

Herrnstein and Murray select 1960 as the year by which America saw potent meritocratic mechanisms in place. This generates a prediction that can be tested against evidence. Recall what a trend toward meritocracy means. The more meritocracy, the more good genes for IQ go to high status occupations, the more bad genes go to low status occupations. The genes are passed on from parent to child, so the more meritocracy, the more of an IQ gap between upper- and lower-class children. If Herrnstein and Murray are correct, the gap between upper- and lower-class children should show a visible jump when we compare representative samples of children tested recently with those tested in the premeritocratic era. The comparability of the most recent data rests on an assumption: that women show no less merit in attaining professional status than men. Social scientists who find life too dull or devoid of controversy are invited to step forward.

The best evidence comes from white American samples, and I have analyzed these to show that they falsify the posited trend toward meritocracy. The correlation between child's IQ and parental occupational status has been surprisingly stable from 1948 to the present. The pattern is a mean IQ of 105 for upper-class children, 100 for middle-class children, 95 for lower-class children. The most parsimonious conclusion is this: nothing, nothing, absolutely nothing has happened.

However, the best that evidence can do is show that meritocratic trends do not exist at a particular time and place. This leaves the central contention of the meritocracy thesis untouched. That contention is that if the humane-egalitarian quest of abolishing inequality and privilege is successful, it will result in class stratification of genes for talent of which IQ is a marker. If such stratification has not occurred, the quest has simply been unsuccessful. Moreover, Herrnstein and Murray claim that a meritocratic future is inevitable. This means that the humane-egalitarian ideal has been given a reprieve both temporary and humiliating. It is a poor ideal that must pray for eternal failure in order to avoid unwelcome consequences. Therefore, we must go beyond evidence to analysis.

The major barrier to abolition of inequality and privilege is our obsession with money and status. Job creation, public health and education, and the welfare state have to be financed by progressive taxation, death duties, luxury taxes. Even limited objectives are costly. One example would be the cost of giving America's depressed urban communities better housing - something which is desirable not only for its own sake but also so that these communities can attract middle-class residents who bring with them their mores and job networks (Dickens, 1999). Which is to say that all of the steps needed to equalize environments involve massive transfers of wealth from some to others. They founder on the rocks of the love of money in one's own pocket, the lust for status superior to one's fellows, the desire to confer advantage for these things on one's family. The fact that universities now do a better job of matching credentials to academic performance does not abolish the enormous inequities of the larger society. Some parents are simply better placed to advantage their children. They provide educationally efficient homes that point children toward superior credentials (Flynn, 1991, 126-139), alter their children's appearance to make them more presentable, give them models of people in work, and pay off crippling debts. Best of all, their contacts and networks become their children's contacts and networks.

Even within the working class, youths can be divided into those who have functional and dysfunctional networks. Wial (1988) describes Boston youth fortunate enough to have fathers and uncles who tell them what skills they need (often learned informally on weekends), what behavior patterns are expected on the job, the importance of avoiding a criminal record, and provide information about job availability. The absolutely crucial role information plays is shown by the fact that about half of all jobs are found through connections (O'Regan, 1993, 329, Table 1). Wial's young men viewed door knocking and answering newspaper ads as equally fruitless. They took it as axiomatic that decent jobs depend on two things only: connections and luck. Youths in families and neighborhoods without viable networks miss out on everything important, no good preparation, no good information, no interview with an employer arranged by a friend working for that employer (Dickens, 1999). Connections and luck are factors whose reach extends right to the top of the job hierarchy (Granovetter, 1974). In 1990 the National Center for Career Strategies stated that over 80 percent of executives find their jobs through networking and that about 86 percent of executive job openings do not appear in the classified advertisements (Ezorsky, 1991, 14-16).

An America in which everyone wants to win the glittering prizes of wealth and status will not pay onerous taxes (or show heroic virtue when tempted to seek special advantage) just so the competition can enjoy a level playing field. However, let us imagine that the value change needed to achieve equal opportunity has occurred: let us imagine what would happen were people to lose their obsession with money and status. The class hierarchy that ranks by income and an agreed pecking order of occupations would be diluted beyond recognition. People must care about that hierarchy for it to be socially significant or even for it to exist. Imagine a society in which the appreciation of beauty, the pursuit of truth, craft skills, being fit, companionship, personal traits like good humor and generosity, and so forth really counted for more than having above average income and possessions. Some people would be better than others at all of these things, but there would be at least a score of noncomparable hierarchies, and being better would not necessarily carry financial rewards. Even today there are executives who care less about promotion than running a good 10 k. The decline of elitist values, less joy in the sheer fact that you are better at something than others are, is also relevant. Superior performance would persist, but less status, less passion, less of a sense of being a better human being would attend superior performance.

In sum: either meritocracy posits a population who are materialist and elitist but who make financial sacrifices and sacrifice the prospects of their children just so others have a better chance to compete, or meritocracy posits today's class system as eternal, even though people have undergone a sea change that has eroded their love of money and status. The present class system cannot become just without a value shift, and a value shift would alter the present class system. Moral realists who believe the last sentence would be improved by calling that value shift a more accurate perception of moral facts are welcome to do so. After all, people have become less "morally depraved".

Meritocracy is also sociologically incoherent: (1) allocating rewards irrespective of merit is a prerequisite for meritocracy, otherwise environment cannot be equalized; (2) allocating rewards according to merit is a prerequisite for meritocracy, otherwise people cannot be stratified by wealth and status (3) therefore, a class-stratified meritocracy is impossible.

This reveals an ambiguity at the heart of the meritocracy thesis, namely failure to specify the quality of the equalized environments

assumed. For most of us, giving everyone equal opportunity would mean everyone with access to quality health care and education; everyone reared in nondemonized homes and communities, that is, by parents in decent housing and with decent jobs; everyone protected against handicaps like having to support a indigent parent or parents. If these things are enjoyed by 95 to 99 percent the population, they can hardly be reserved to those of outstanding merit.

Yet equalization of environments is to coexist with a large immiserated underclass, and that class must compete with an elite that has an environment so potent that they constitute a menace to democracy (Herrnstein & Murray, 1994,509-526). The ideal that truly self-destructs in practice is the meritocratic ideal. Those who think it inevitable should give it a plausible social dynamic. They can begin by telling us how equality is to be achieved when a large underclass is already knocking at the door, or, conversely, how an underclass is to emerge if we keep topping up their environmental quality to maintain the level needed for equal opportunity. It is significant that Herrnstein and Murray imagine environments being equalized by magic. Magic's next task is to reconcile equality with a large underclass. Its final task should be to square the circle.

Our sociological analysis reinforces our psychological analysis. The higher we push the quality of environment all enjoy, the less attractive the prizes left for the winners. Many people of talent may want more than the not-unattractive norm, but how many will care about shaking the last dollar out of the money tree? Social scientists can go on publishing hierarchies that rank the whole population by occupational status, but these will fall short of ranking people by merit, much less genes for talent. An overenthusiastic sports master can force everyone to participate in the annual school run, but he or she cannot force them to train or try. The published results will not stratify people for genes for running ability. A decent life for all does not foster a social Darwinist psychology or raise competition to fever pitch.

Now we have a better understanding of the dynamics of humane egalitarian ideals. Rather than self-destructing in practice, they possess a self-correcting mechanism that avoids meritocratic excess. The truth is that we cannot push equality much beyond our ability to humanize. Every significant step toward equality must be accompanied by the evolution of values unfriendly to "success" as defined by the present class structure. Every significant step toward equality means a step toward a people less materialistic and elitist, more variegated in their interests and behavior,

altogether more humane. Whatever dark spirits lurk in the depths of equality, meritocracy is not among them.

A final disclaimer: this analysis makes no prediction about how far we can go toward humanizing people away from materialistic and elitist values: it does not even say how far we should go. The caution does not come from recognizing that people disadvantage blacks because of rational self-interest. The fact that bankers, landlords, employers, and proprietors want to survive market competition is quite compatible with putting your woodworking hobby ahead of plotting to be president of General Motors. The caution comes from an inability to predict history. What the analysis does attempt is to describe the interaction between humane values and egalitarian ideals, to show that radical progress beyond the status quo for one assumes radical progress for the other. It attempts to show that when our critics write a scenario that assumes radical equality of opportunity conjoined with the present class system and its psychology, they simply are not thinking clearly.

References

Ayers, I., & Siegelman, P. (1995). Race and gender discrimination in bargaining for a new car. *American Economic Review, 85*, 304-321.

Berger, J. (1987). New York Times News Service, June 19.

Bertrand, M., & Mullainathan, S. (2003). Are Emily and Greg more employable than Lakisha and Jamal? *A Field Experiment on Labor Market Discrimination*. National Bureau of Economic Research, NBER Working Paper W 9873 (July 2003).

Dickens, W. T. (1999). Rebuilding urban labor markets. In R. F. Ferguson & W. T. Dickens (Eds.), *Urban Problems and Community Development*. Washington DC: Brookings Institutional Press.

Ezorsky, G. (1991). *Racism and Justice: The Case for Affirmative Action*. Ithaca NY: Cornell University Press.

Flynn, J. R. (1991). *Asian Americans: Achievement Beyond IQ*. Hillside, NJ: Lawrence Erlbaum Associates.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54(1)*, 5-20.

Granovetter, E. M. (1974). *Getting a Job: A Study of Contacts and Careers*. Cambridge MA: Harvard University Press.

Herrnstein, R. J., & Murray, C. (1994). *The Bell Curve*. London: MacMillan.

Hindelang, M. J. (1978). Race and involvement in common law personal crimes. *American Sociological Review, 43*, 91-109.

Levin, M. (c.1991, unpublished). *Responses to Race Differences in Crime*.

Maddox, K. B., & Gray, S. A. (2002). Cognitive representations of black Americans: Reexploring the role of skin tone. *Personality and Social Psychology Bulletin, 28*, 250-259.

O'Regan, K. M. (1993). The effect of social networks and concentrated poverty on black and Hispanic youth unemployment. *Annals of Regional Science, 27*, 327-342.

Rushton, J. P. (1988). Race differences in behaviour: A review and evolutionary analysis. *Personality and Individual Differences, 9,* 1009-1024.

Stove, D. (1995). *Cricket Versus Republicanism and Other Essays*. Sydney: Quakers Hill Press.

U.S. News and World Report. (1998). The black-on-black crime plague.

Wail, H. (1988). *The transition from secondary to primary employment: Jobs and workers in ethnic neighborhood labor markets*. Ph.D. dissertation, Department of Economics, MIT, Cambridge, MA.

Wilson, W. E., & Herrnstein, R. J. (1985). *Crime and Human Nature*. New York: Basic Books.

Chapter 26

Social Cage (socio-economic status and intelligence in Hungary)*

Balázs Klein, Sándor Klein, Kálmán Joubert, and Gyula Gyenis

These days we need to explain to our children what the historical notion "conscription" meant. Not too many people feel sorry about the abolition of the compulsory military service in Hungary. But it had at least one desirable side effect: It created the opportunity to test a very large and representative cross section of the population within a well defined age-group.[26.1] In this article we report some results from one such study.

Out of the 73 thousand conscripts in 1998, a representative sample of 8000 18 year old male conscripts was selected by the Hungarian Central Statistical Office (HCSO)[26.2]. Almost 7000 of them completed the Standard Progressive Matrices ***Plus*** (SPM+) and a questionnaire covering many background variables. This chapter explores the interesting relationships between these variables themselves and with SPM+ scores.

Our working definition of socio-economic status (SES)

Because very few of the conscripts answered the direct questions about the incomes of their families - which could have led to a more sophisticated classification of their background socio-economic status - it was necessary to construct an index based on the *Combined Educational Level of their Parents* (CELP).

* A short version of this chapter was published in Hungarian: Klein, B. – Klein, S. – Joubert, K. – Gyenis, Gy.: Intelligencia és iskolázottság Magyarországon (Intelligence and schooling in Hungary) *Mozgó Világ,* 2006, June. The study was designed by Joubert, K. and Gyenis, Gy. This report and its earlier version were prepared by Klein, B. and Klein, S.

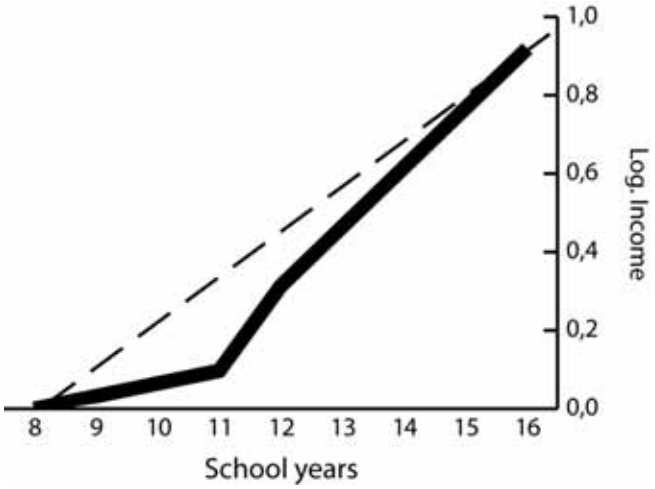
There is ample evidence from other studies that education and income are strongly correlated. Kézdi (2005), based on Hungarian national statistics, shows that this is also true in Hungary (Figure 26.1).[26.3] The Hungarian national statistics also show that, not only is there a strong correlation between education and income, but that the income yield of an additional school year to the "standard" 8 years of schooling steadily increased between 1989 and 2002, thus reinforcing the common belief that knowledge – or at least credentials – is/are getting more and more important in the labour market (Figure 26.2).[26.4]

We calculated our CELP (Combined Educational Level of Parents) indices by summing the educational levels of both parents (coded individually) according to the following scheme (cases where data was not given for one or both parents were omitted from the calculations):

- 0 points - unfinished elementary school
- 1 point – finished elementary school
- 2 points – secondary school
- 3 points – higher education

The published statistical data about the education of the Hungarian people present a depressing picture: Almost a million adults did not

Figure 26.1. **Income Yield of Schooling in Hungary***
Hungarian Statistical Bureau, 2002



* Extracted from Gábor Kézdi (2005), Education and Earnings pp. 31-37 *in* Károly Fazekas and Júlia Varga (eds.), *The Hungarian Labour Market Review and Analysis, 2005*. Budapest: Institute of Economics HAS - Hungarian Employment Foundation.

Figure 26.2. **Income Yield of an Additional School Year to the Standard Eight Years of Schooling (in Percentages)**
Hungarian Statistical Bureau, 2002

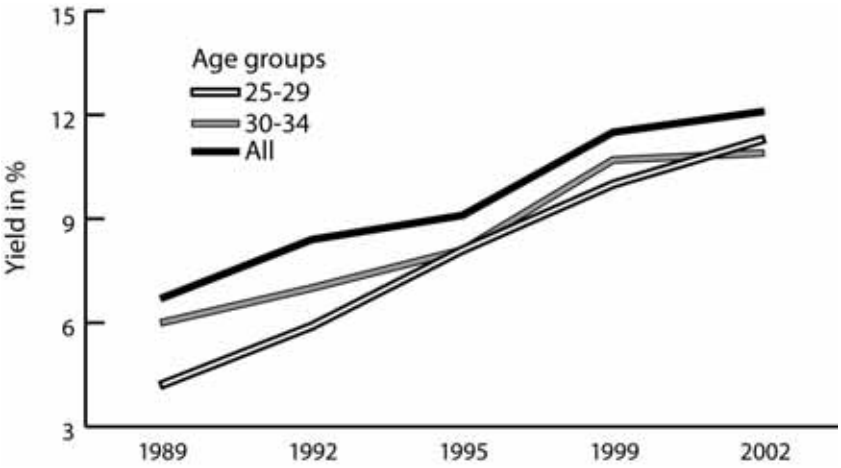
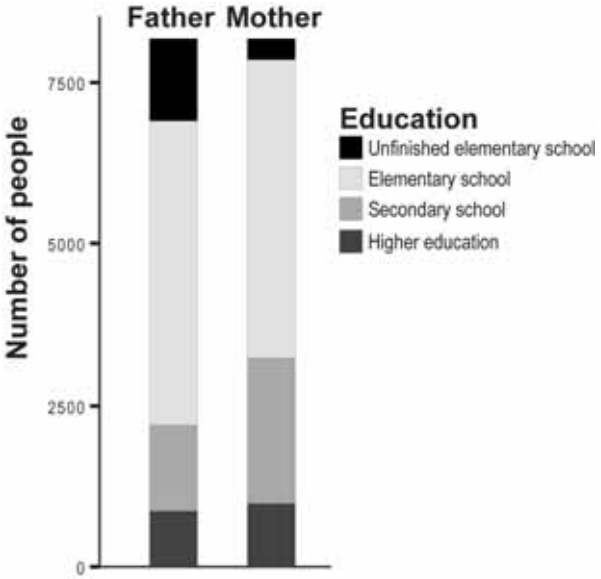


Figure 26.3. **Education of Parents**
1998 Sample of Conscripts

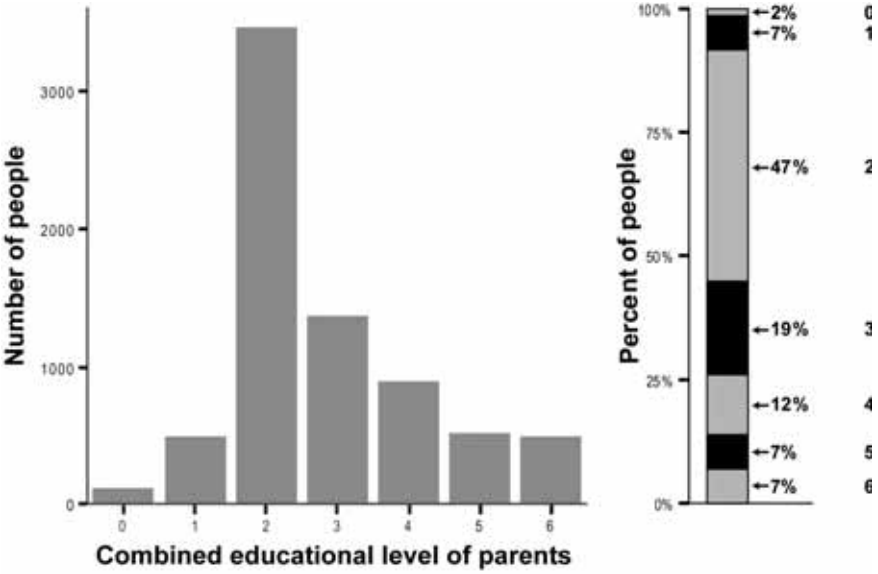


complete elementary education (and males were less likely to do so than females). These overall statistics are well reflected in our data (Figure 26.3).

In what follows we will use our Combined Educational Level of Parents (CELP) indices as a proxy for the 18 year olds' socio-economic status (SES).

Looking at the distribution of the CELP (Figure 26.4) we see a distribution skewed to the left where most families would have a combined educational level of 2. As we can see this is mostly as a result of both parents having only an elementary school education.

Figure 26.4. **Distribution of the Combined Educational Level of Parents**
1998 Sample of Conscripts



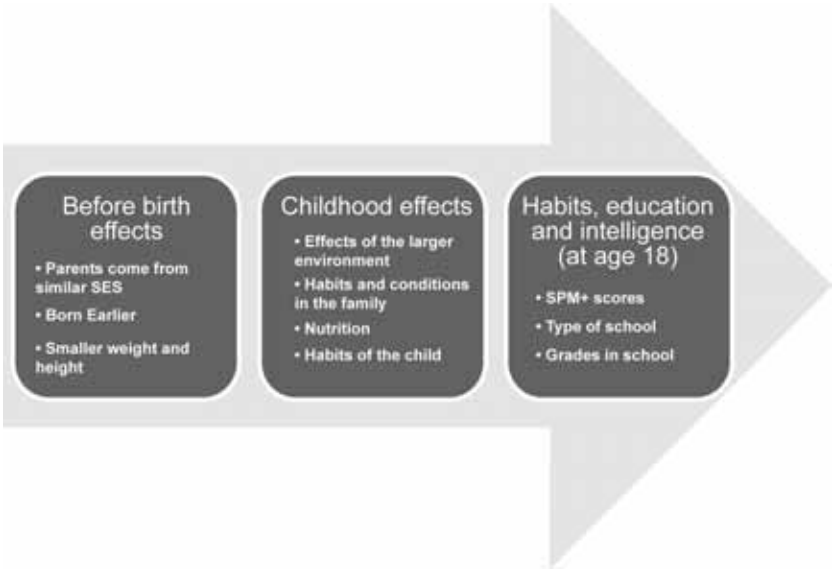
The re-generation of socio-economic status in society

In this part of the paper we explore how SES regenerates itself. We explore this issue in three stages (Figure 26.5)

1. **Before birth effects**
   Parents tend to marry from the same SES than themselves. The environment and habits of parents have a significant effect on the attributes of their child at the earliest age.
2. **Childhood effects**
   Through the environment and habits of the family - and increasingly through the personal choices of the son himself - the effects of SES are amplified during the years of childhood and education.
3. **Habits, education and intelligence (at the age of 18)**
   By the age of 18, the difference between the SPM+ scores of young people from families having different CELP indices is significantly different not only statistically but also in absolute terms: CELP – and so probably SES – largely determines the level of education of the young generation in Hungary (see later).

We will consider these three stages in more detail now.

Figure 26.5. **Steps to Regenerate Socio-Economic Status**



1. Before-birth effects

There are many indications that SES affects the future of the unborn babies through the nutrition, type of work, habits (like smoking, drinking etc.) of parents (especially the mother). After demonstrating how similar parents' educational levels in Hungary are, we will see that low SES children are born sooner, with smaller height and weight.

We seem to live in social cages – marriages are most likely made among people with very similar educational levels. The correlation between the educational levels of parents was 0.57. The difference between the educational levels of parents is most likely to be zero (Figure 26.6). In fact two-thirds of the parents in our study had identical educational levels (mostly they both had only elementary education; Figure 26.7).

In families with very low socio-economic status (CELP 0 or 1) children are born significantly sooner (Figure 26.8) with lower height (Figure 26.9)

Figure 26.6. **Difference between Educational Levels of Parents**
1998 Sample of Conscripts

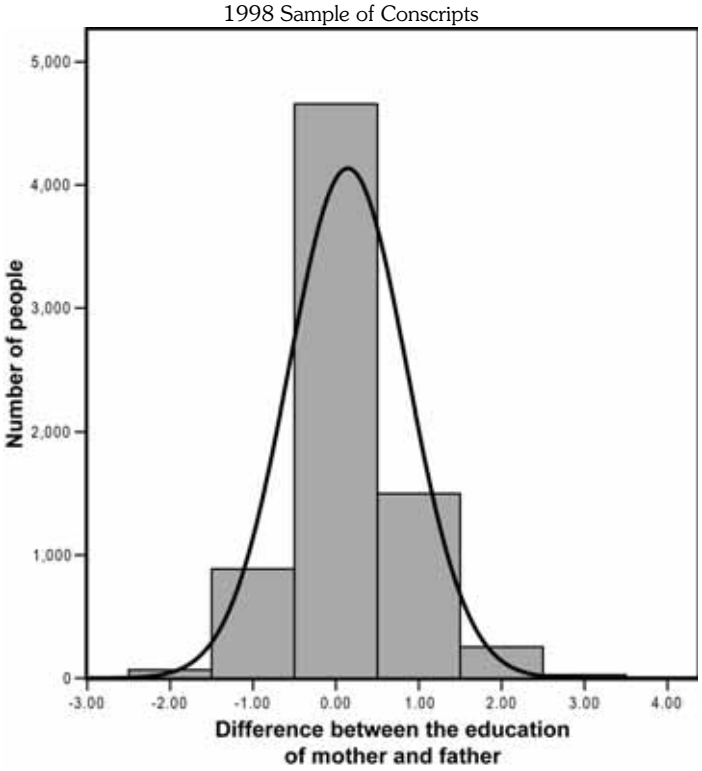


Figure 26.7 **Educational Level of Parents**
1998 Sample of Conscripts

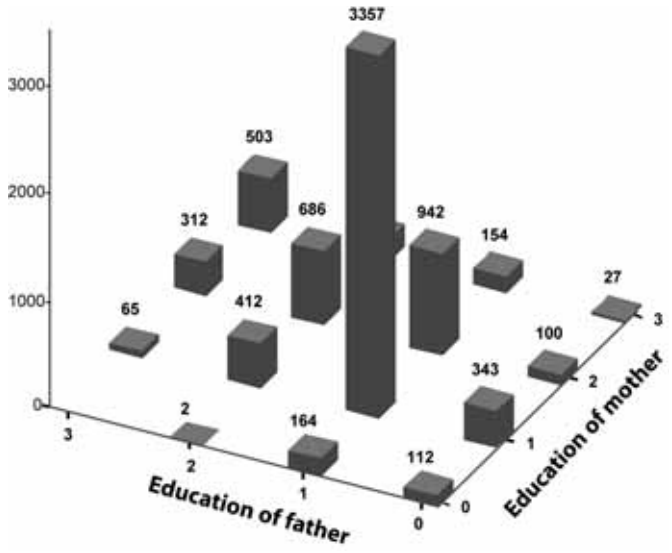


Figure 26.8. **Week of Birth in Relation to the Combined Education Level of Parents**

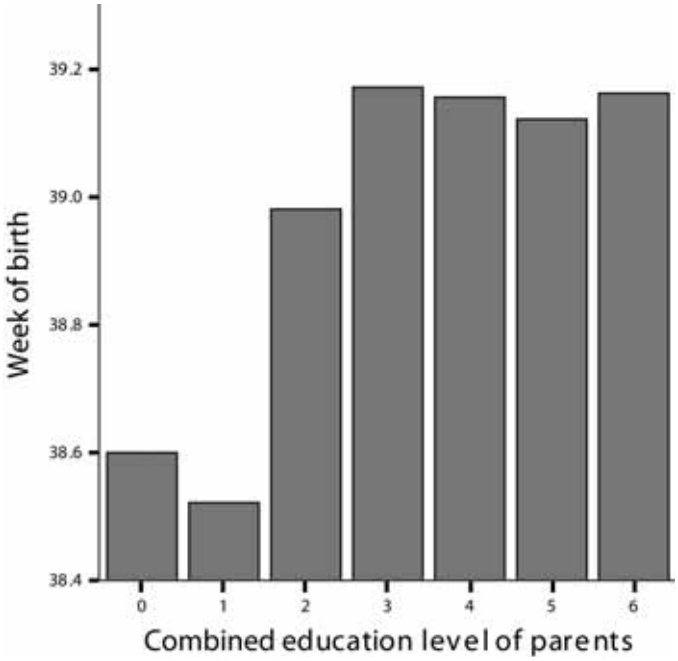


Figure 26.9. **Height at Birth in Relation to the Combined Education Level of Parents**
1998 Sample of Conscripts

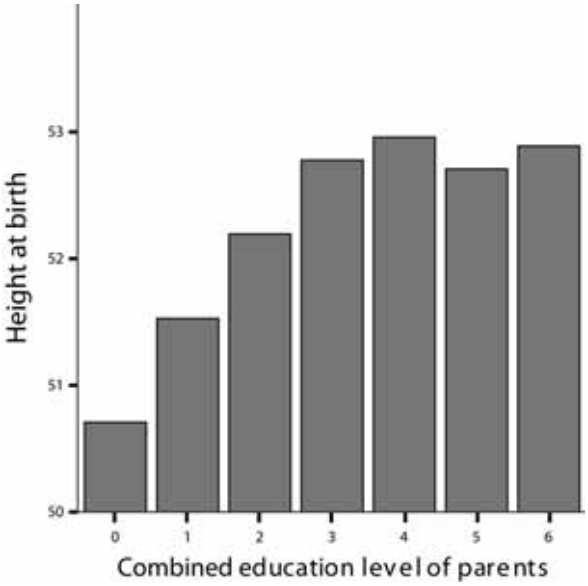
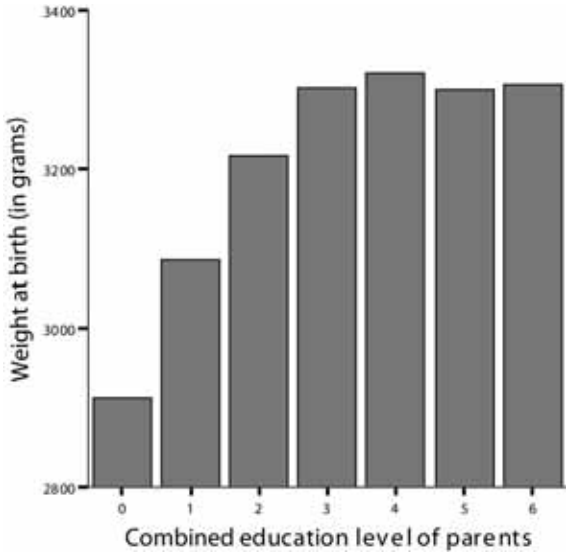


Figure 26.10. **Birth Weight in Relation to the Combined Educational Level of Parents**
1998 Sample of Conscripts

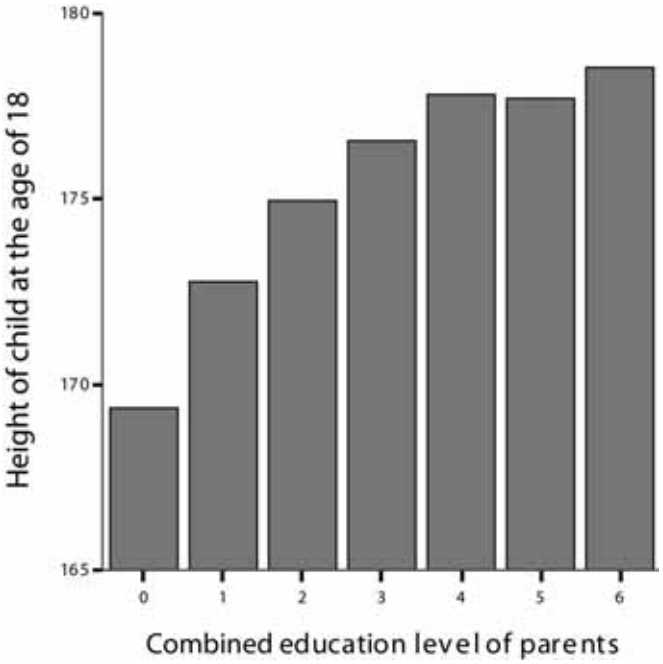


and weight (Figure 26.10) than those raised in higher socio-economic status groups (where CELP is 3 or more). In the later group of families there is no detectable relation between socio-economic status on one side and height or weight at birth or length of pregnancy on the other side. The biggest group – 47% of the families, where CELP is 2 – lies in these respects between the lower (0-1) and higher (3-5) CELP groups.

The relationships documented in Figures 26.8, 26.9 and 26.10[26.5] may be due to such things as the lifestyles of the parents including their nutrition and general health prior to giving birth. Or they may be, at least in part, due to genetic characteristics.

At the age of 18, the relationship between the socio-economic status of the parents and the respondent's height is even stronger then it was at birth (Figure 26.11).

Figure 26.11. **Height of Son at Age 18 in Relation to CELP**
1998 Sample of Conscripts



2. Childhood effects

In this section we enumerate a number of factors that – through the environment, and habits of the family, but also through the own choices of the child – amplify the effects of socio-economic status:

- Area of Residence of the respondent
- Number of books in the family
- Having a computer in the family
- Nutritional habits
- Other habits like smoking, number of times the respondent takes a shower or brushes his teeth.

Rather than suggesting that these variables are all in a causal relation to the educational level of the parents, they point to the conclusion that there is a very wide range of interdependent circumstances and behaviours that recursively strengthen the effect of each other and are all strongly related to CELP, and so possibly socio-economic status.

Area of residence

Figure 26.12 shows the percentages of respondents residing in different types of settlement. This is a good sample of the total Hungarian population. In both the sample and the nation, most of the population live in smaller settlements.

Figure 26.12. **Distribution of the Residence of the 1988 Sample of Conscripts**

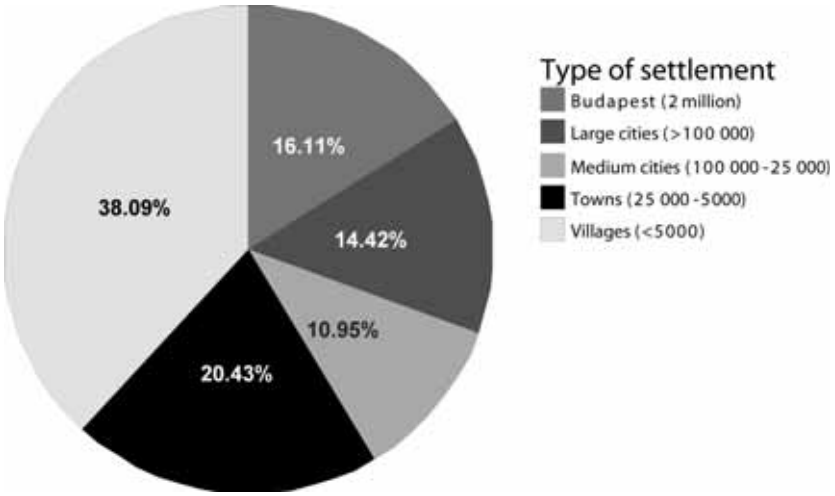
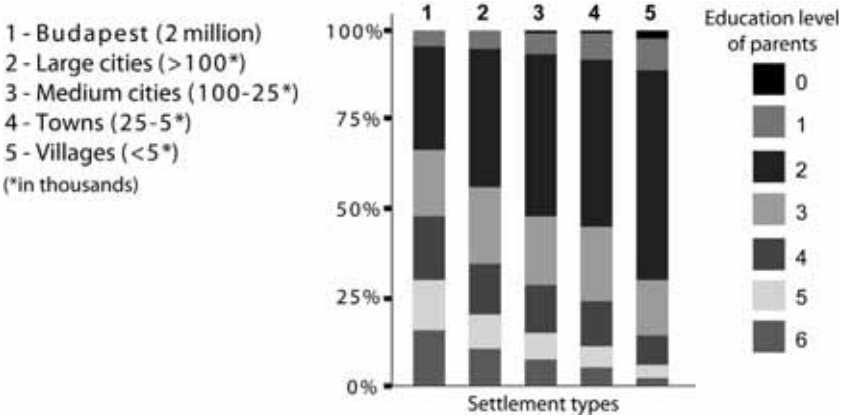


Figure 26.13. **Education Level Of Parents in Different Settlement Types**
1998 Sample of Conscripts



The educational level of the parents varies between the types of settlement (Figure 26.13): respondents who have parents having higher educational levels tend to live in larger settlements.

Number of books in the family

Many studies have shown that the relationship between children's school performance and the "cultural goods" in the family is even stronger than that with "monetary goods".

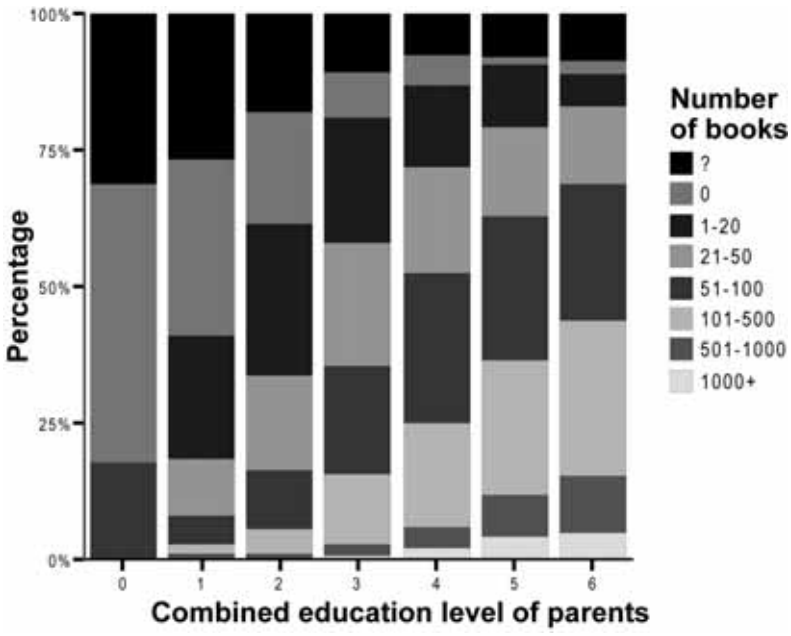
Figure 26.14 illustrates that, in our sample, the number of books in the family increases steadily with the combined educational level of the parents. (The category with the question mark is made up of respondents who – due to ambiguity in the instructions – entered 0 as the category number. Since this category did not exist in the questionnaire we believe that – like those in the next category – these respondents were indicating that there were no books in their homes.)

No wonder that in Hungary the reading achievement of children from high SES families is far better than that of children from low SES families – and that this difference is much bigger than in most other countries.[26.6]

Having a computer in the family

Computer literacy can be a passport to many future jobs and opportunities. Having a computer in the family probably increases the chances that a

Figure 26.14. **Number of Books in the Family in Relation to the Combined Education Level of the Parents**
1998 Sample of Conscripts



young person will have a useful knowledge of computers by the time he leaves school. As might be expected, the chances of finding a computer in a household depends heavily on the socio-economic status of the family (Figure 26.15).

Nutritional habits

Respondents indicated the frequency with which they consumed 19 different types of food. Out of these data we created nutrition habit factors by studying the relationships between the consumptions of these products. We named the three nutritional habit factors we identified as Fast food, Healthy food and Fat food (Table 26.1).

As expected, nutritional habits vary between respondents coming from different socio-economic backgrounds (Figure 26.16):

- there is a positive correlation between SES and the frequency of Healthy food consumption – respondents from low Socio-Economic backgrounds eat little Healthy food, while those from high SES backgrounds eat significantly more,

Figure 26.15. **Having a Computer in the Family Depends on CELP**
1998 Sample of Conscripts

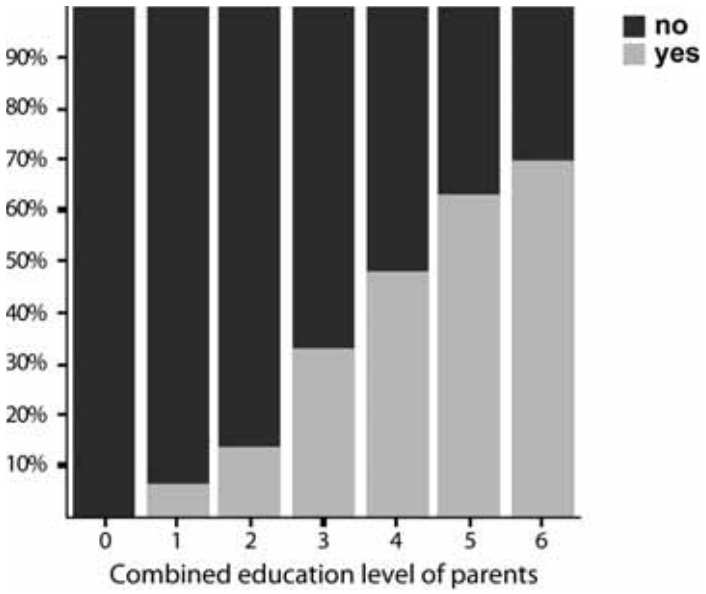
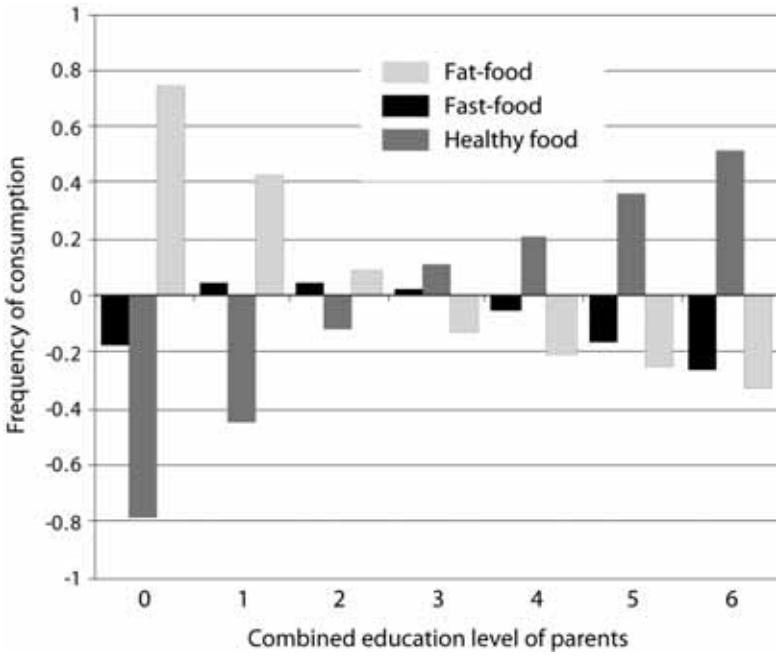


Table 26.1. Factor Loadings on the Three Factors of Different Types of Food Consumption (only absolute values larger than 0.15 appear)

	Fast food	Healthy food	Fat food
Hamburger, Hot-Dog	.80		
Cola	.70		
Chips	.76		
Vegetables (raw)		.76	
Fruits		.74	
Milk, cheese, quark		.67	
Bacon			.82
Bread and dripping			.83

Figure 26.16. **Nutrition Habits and Socio-Economic Status**
1998 Sample of Conscripts



- there is a negative correlation between SES and the frequency of Fat-food consumption – respondents from low SES backgrounds eat a lot of Fat-food, while those from high socio-economic backgrounds eat less,
- respondents from both very low and very high SES backgrounds eat Fast food more frequently than those in between.

3. Habits, education, and intelligence (at age 18)

HABITS

Childhood environment can have a long-lasting effect.[26.7] Below we will see that the habits and attitudes of our 18 year olds were closely related to the SES of their families (in the following low SES = CELP 0-2, high SES = CELP 3-6).

Drinking

Drunkenness is less common among 18 year olds coming from low CELP families than those from high CELP families (Figure 26.17).

Drugs

Although those from High SES backgrounds were more likely to disapprove of trying cannabis (Figure 26.18), there was no difference in the frequency of regular use.

Figure 26.17. **Socio-Economic Status and Overdrinking (i.e. Have you been drunk in the last month?)**
1998 Sample of Conscripts

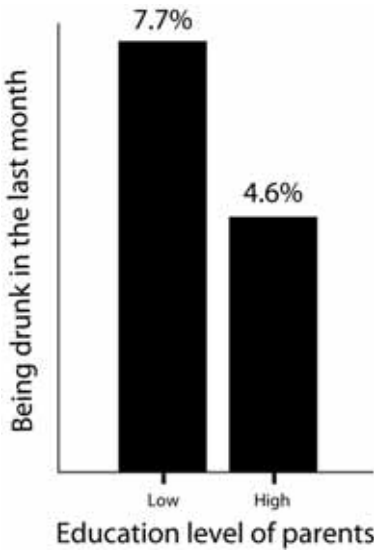


Figure 26.18. **Socio-Economic Status and Drugs (i.e. Do you disapprove trying cannabis?)**
1998 Sample of Conscripts

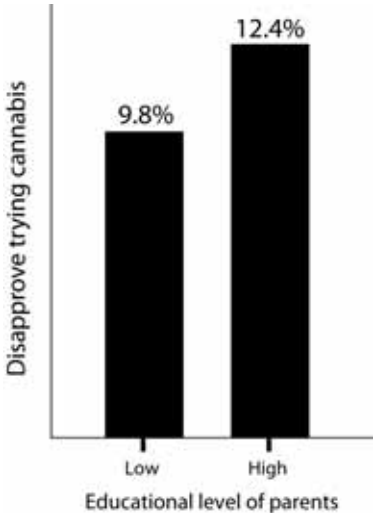
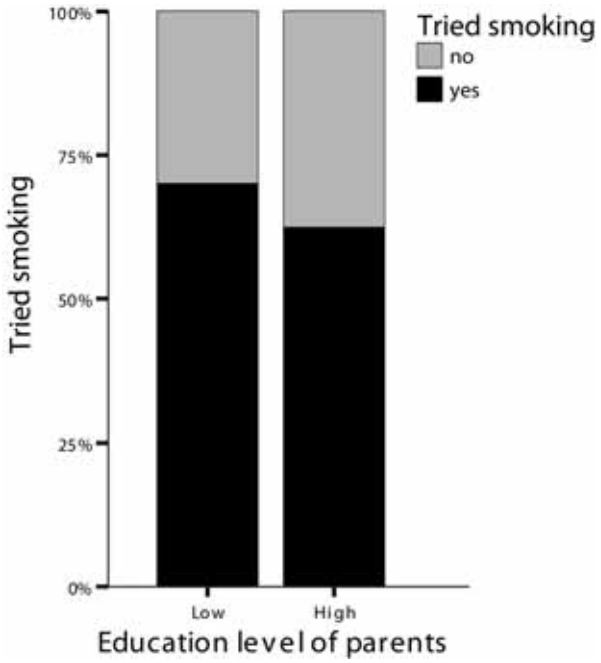


Figure 26.19. **Smoking and Socio-Economic Status (i.e. Have you ever tried smoking?)**
1998 Sample of Conscripts



Smoking

Low SES background is somewhat associated with smoking.

EDUCATION

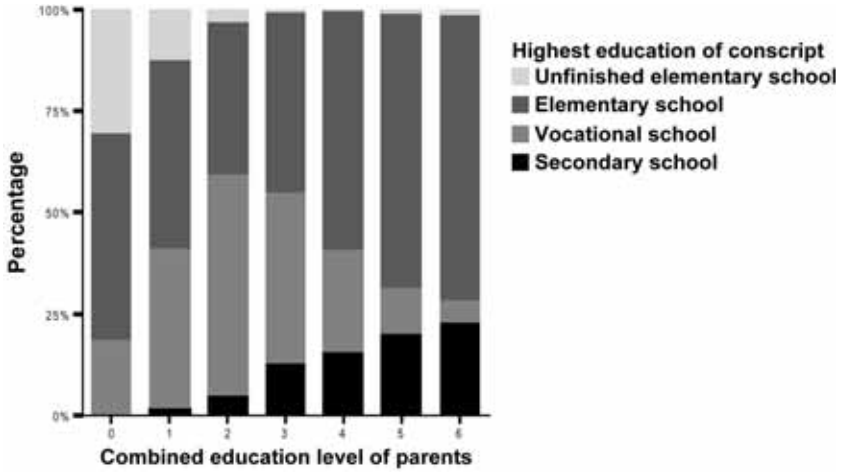
The *highest* education the conscript obtained by the time of the conscription at the age of 18 was strongly related to CELP (Figure 26.20).

Practically the only respondents who could not finish elementary school by the time of conscription (18 years of age) were those whose parents' combined educational level was 2 or less.

The highest proportion having attended vocational school were those whose parents had a combined educational score of 2. Thus it seems that vocational school is possibly too demanding for youngsters coming from 0 and 1 backgrounds but is less and less attractive to those whose parents had combined scores of 3+.

The percentage of respondents who had completed secondary school by the time they were 18 steadily increases with the combined

Figure 26.20. **The Highest Education of the Conscript in Relation to the Combined Educational Level of the Parents**
1998 Sample of Conscripts



educational level of their parents. Where only one parent had finished elementary school the respondent had a less than a tenth of the chance (2% as opposed to 23%) of getting into secondary school than those whose parents both had higher education. In fact, we were shocked to find that, in families where neither of the parents had finished elementary school (more than 100 cases) not one of the respondents had completed secondary school – which is virtually the only way to get into university – by the age of 18.

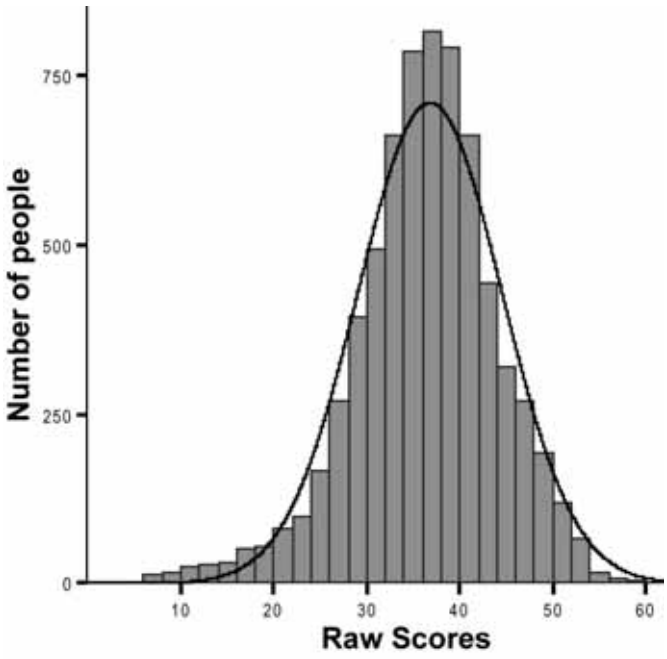
INTELLIGENCE

Distribution and validity of the *Raven Standard Progressive Matrices Plus* data.

Although the lower tail of the distribution was more curtailed than we expected, there was no major distortion in the distribution and the average score of the sample was exactly the same as in the representative sample of 18 year olds in Romania (Figure 26.21).

To compensate for intentional faking of low scores (see the chapters in this book authored by Kim McKinzey), we omitted all scores that were more than 4 standard deviations below the mean. In practice this meant a raw score of 6 or below. Even by answering at random a respondent would normally get a score of about 7. By doing this we excluded 0.2

Figure 26.21. *Standard Progressive Matrices Plus*
Hungarian Standardisation among 18 year Old Conscripts
Distribution of raw scores



percent of our sample. Altogether we believe that this exclusion did not significantly change the properties of the sample.

The average of the resulting distribution was 36.8; its standard deviation was 7.7. The maximum score in the sample was 60 – the highest possible score in the test. Scores below 6 were treated as missing data.

Socio-Economic Status and SPM+ score

A large body of literature exists on the relationship between SES and RPM scores.

However, we must begin with a warning regarding overinterpretation of differences between groups. It must always be borne in mind that there are huge individual differences within each group.

More than half a century ago, Anastasi noted in his *Differential Psychology* that, since the assessment of the American soldiers' intelligence during the first world war, the relationship between intelligence

and SES is the best documented fact in psychology.[26.8] In the 40's a study with 80 thousand respondents demonstrated that there are significant differences in the intelligence of people having different jobs (e.g. accountants, teachers vs. workers, farmers). In America, also in the 40's, the intelligence of kindergarten age children coming from highly educated families were 20 points (more than 1 SD) higher than those whose parents were workers, and this difference was about the same at age 18.

We will not enter here into the debate about heredity vs. environmental effects[26.9] but simply want to state that large-scale Hungarian studies – being part of international evaluations – show that in our country the achievement differences in cognitive tasks (e.g. tasks with inductive thinking, mathematical problems) between children coming from different SES families are especially big.[26.10] According to the PISA – 2000 study[26.11], carried out in 32 countries with 250 thousand students, the students coming from first quartile SES families have twice as big chance to have their school achievements in the lowest 25 percentile, than their schoolmates. And among all the countries in this study the achievement differences attributed to SES were the biggest in Hungary.

In the course of the present study we have already mentioned several differences between different SES groups, but none of these were as profound as the differences in intelligence. Figure 26.22 shows how the average score on SPM+ increases with the combined education level of parents.

While in families in which neither of the parents completed elementary school respondents' average score on the SPM+ was 28, in families where both parents had higher education, the average score was 43 – 2 Standard Deviations higher.

In fact, as Figure 26.23 shows, there is very little overlap between the scores of respondents from the highest and lowest educational backgrounds.

We can examine the relationship between SPM+ scores and combined educational level of parents' (shown in Fig. 26.22) in more detail by looking at the effect that the educational level of each parent separately has on SPM+ performance (Figure 26.24).

Figure 26.24 shows that SPM+ scores increase with the educational level of either parent. However, the most dramatic increase occurs among those whose father did not complete elementary school. It is also interesting to note that, so far as SPM+ score is concerned, those whose parents have a university or college education have only a small advantage over those whose parents have only secondary education.

Figure 26.22. *Standard Progressive Matrices Plus*
SPM+ Scores as a Function of Parental Education
Hungarian Standardisation among 18 year Old Conscripts

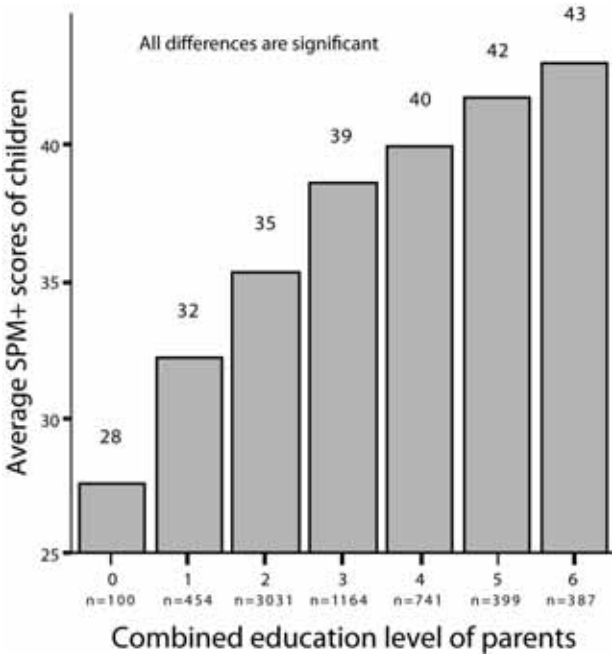


Figure 26.23. *Standard Progressive Matrices Plus*
SPM+ Scores of Conscripts with Very Low and Very High Socio-Economical Background
Hungarian standardisation among 18 year old conscripts

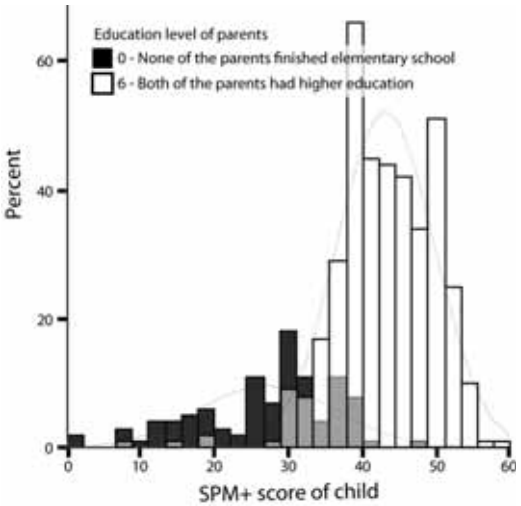
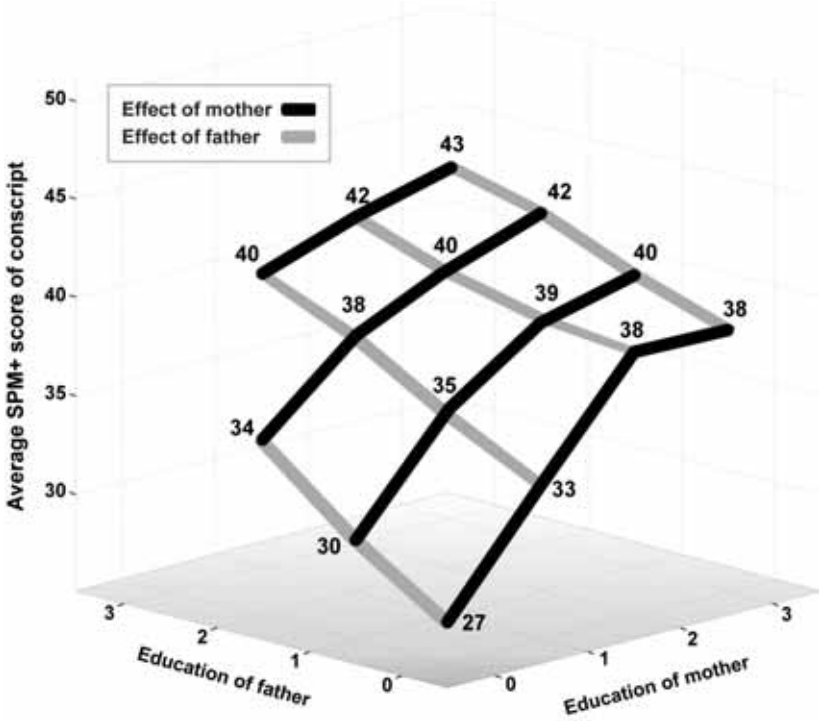


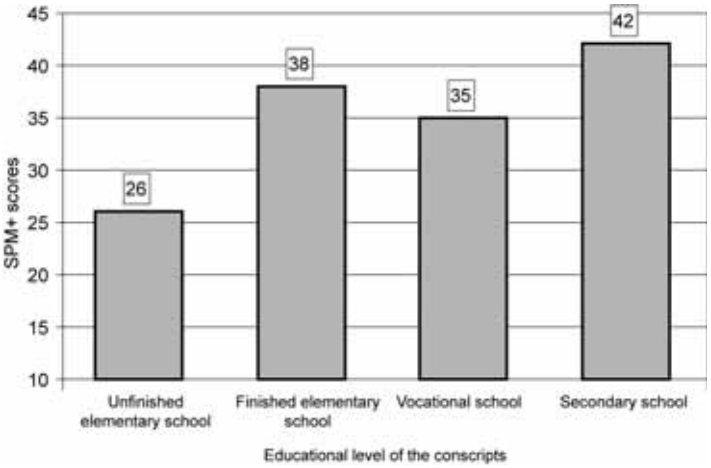
Figure 26.24. *Standard Progressive Matrices Plus*
Scores of Conscripts in Relation to the Educational Levels of Both Parents
Hungarian standardisation among 18 year old conscripts



Schooling and intelligence

Many studies have shown that educated people are more intelligent, so it is not surprising that, in our study, those who finished elementary school have higher SPM+ scores than those who did not and that those who finished secondary school have, on average, higher scores than those who only finished elementary school. More surprising is that those who attended a two year vocational school had average scores lower than of those who did not study anything after elementary school (Figure 26.25). Of course we do not know whether only young people of lower ability selected vocational school or if this type of school has a bad effect on the cognitive ability of the students. But, in either case, it is bad news for a country which badly needs able people in these important vocations.

Figure 26.25. *Standard Progressive Matrices Plus*
Distribution of Scores According to the Educational Level of Conscript
Hungarian standardisation among 18 year old conscripts

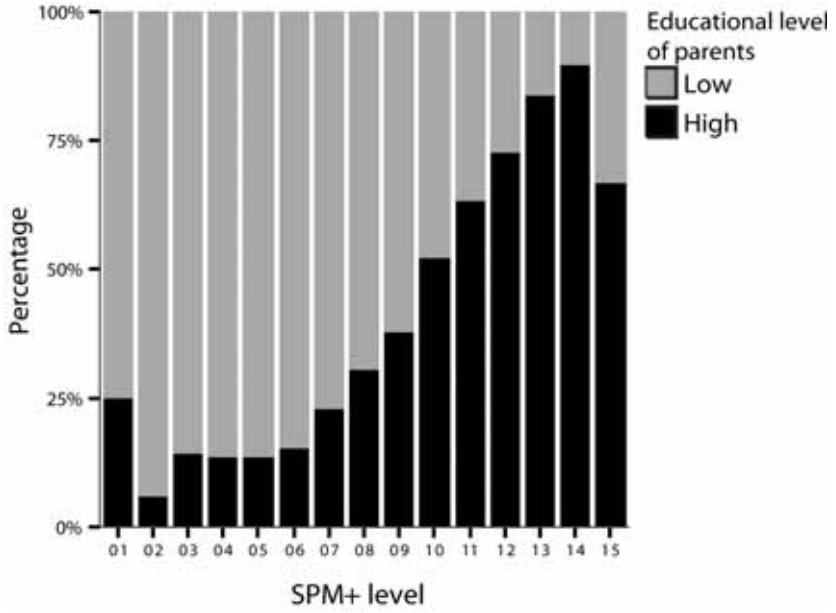


Intentional distortion

In reality, the relationship between socio-economic status and RPM score is probably even more pronounced than would seem to be the case from because, as we will now show, there is an indication that some of those from higher socio-economic backgrounds deliberately faked low SPM+ scores.

To investigate this possibility we categorised the raw scores into 15 categories ranging from a mean of 7 to 58. In each of these 15 groups we calculated the ratio of families from high and low socio-economic status backgrounds. As can be seen from Fig. 26.26, the percentage of respondents coming from families with high socio-economic status steadily increases with the SPM+ performance of the son – except in the two extreme categories. There is a much lower proportion of respondents from high socio-economic status backgrounds among those who performed extremely well on SPM+ than there "should" be. And far more respondents from well situated families get poor SPM+ results than would be predicted from the rest of the distribution. Our explanation of these discrepancies is that a number of respondents from high socio-economic status background deliberately faked low scores. (We may note in passing, however, that this distortion probably has little effect on the overall results reported in this chapter since the number of respondents in

Figure 26.26. *Standard Progressive Matrices Plus*
Ratio of High/Low Socio-Economic Background Among Conscripts with Different Performance on SPM+
Hungarian standardisation among 18 year old conscripts



the two extreme groups put together is less than 0.5 percent of the total sample, i.e. 24 people altogether.)

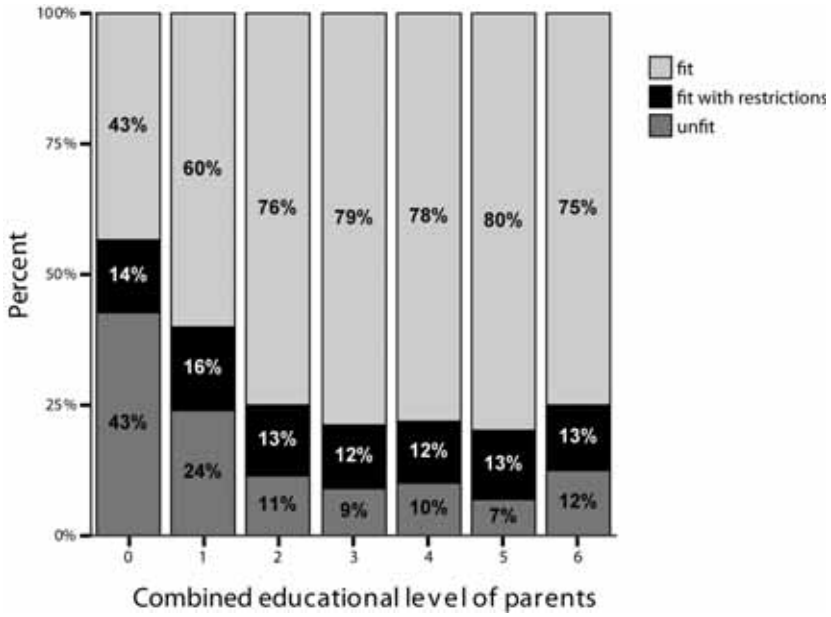
These faked low scores stem from the commonly held belief that an extremely low intelligence test score could lead to excusal from military service (as from the death penalty for murder in the US Chapter 24).

Fitness for military service

It is interesting to note as an aside that there is a significant relationship between socio-economic status and fitness for military service. Respondents with very low socio-economic status proved to be significantly less fit for military service than the others.

As we have seen, socio-economic status is closely related to some sort of mental fitness as measured by the SPM+. The fact that this increase in mental fitness does not translate into fitness for military service (it generally stays the same above level 1) suggests that some respondents from higher socio-economic status backgrounds successfully avoided being drafted into the army (Figure 26.27).

Figure 26.27. **Fitness for Military Service as a Function of the Combined Educational Level of the Parents**
1998 Sample of Conscripts



Concluding remarks

Looking at the various data gathered in 1998 from a representative sample of Hungarian male 18 years olds as a whole, we can see that the effects of differences between the socioeconomic status of the parents (at least as indicated by combined educational level) show up in all manner of ways ranging from birth height and weight to mental capacity and schooling.

Several earlier studies have shown that parents' SES determines the future of Hungarian children much more strongly than is the case in many other countries. So it seems urgent to provide those young people who came from families of lower socio-economic status with a fair chance to develop their full potentials. Schools have a responsibility in this task, but they cannot solve it alone (although it is also true that most of them don't even try to work toward this goal). It would seem that, in Hungary, which became a "market economy" less than 20 years ago, the rich are getting richer and the poor poorer whatever the lip-service politicians pay to "social equality".

We do not have a solution at hand, but we do think that every individual can do something. Every individual is a unique world, with a special life-history, personality, knowledge and ability pattern. Environment affects what we will become, but personal freedom and responsibility also have a crucial significance. It is difficult to break out from the cage of SES. In order to help people to realize their full potential, we have to work toward a society which will provide a humane environment for its members, a country where everybody has a fair chance to get proper schooling, where family life is much more then a collective struggle for life, and where schools are places where people go because they learn there something significant, and because they like to go there.

Notes

1. It may be noted that it was precisely because the military systems of many countries had tested birth cohort after birth cohort of young people entering military service on the RPM over very many years that Flynn (1987) was able to assemble the data that enabled him to demonstrate so convincingly the international intergenerational increase in scores that now bears his name.

   Flynn, J. R. (1987). Massive gains in 14 nations: What IQ tests really measure. *Psych. Bulletin,* 101, 171-191.

2. Joubert, K., Gyenis, Gy. (2001): *State of Health and Physical Development of 18 year old conscripts I.* Kutatási jelentések (Research Reports) 70., Népességtudományi Intézet (Demographic Research Institute), Budapest.

3. Károly Fazekas, Júlia Varga (eds.) (2005): *The Hungarian Labour Market Review and Analysis,* Institute of Economics HAS - Hungarian Enployment Foundation. Gábor Kézdi: Education and Earnings pp 31-37.

4. See also: Ceci, S.J., Williams, W.M. (1997): Schooling, intelligence and income. *The American Psychologist,* 52, 1051-1058.

5. Joubert, K. (1983): *Birth weight and birth length standards on basis of data of infants born alive in 1973-78.* Research Reports of the Demographic Research Institute, 12.

   Joubert, K., Darvay, S., Ágfalvi, R. (1996): Growth and development curves for a nation-wide longitudinal growth study of Hungarian children. In: Bodzsár, B. É., Susanne, C. (Eds.): *Studies in Human Biology.* Eötvös University Press, Budapest, 147-156.

Gyenis, Gy., Joubert, K., Klein, S., Klein, B. (2004): Relationship among body developments, socio-economic factors and mental abilities in Hungarian Conscripts. Anthrop. Közl., 45, 165-172.

6. Vári, P. (ed.): *PISA-vizsgálat 2000* (The PISA-2000 study). Műszaki Könyvkiadó, Budapest, 2003, pp. 146.

7. Gyenis, Gy., Joubert, K.(2002): Some characteristics of the health status of the 18-year-old conscripts in Hungary. *Humanbiologia Budapestinensis.* Budapest, 27, 95-105.

   Susanne, C. (1980): Socioeconomic differences in growth patterns. In: Johnston, F. E., Roche, A. F., Susanne, C. (Eds.): *Human physical growth and maturation.* Plenum Press, New York – London, 329-388.

   Tanner, J. M. (1990): Growth as a mirror of conditions in society. In: Lindgren, G. W.(ed.): *Growth as a mirror of conditions in society.* Stockholm, Institute of Education Press, 9-48.

8. Anastasi, A. (1958): *Differential Psychology.* MacMillan, pp. 512-520. The following examples are also from this book with proper referencies to the studies.

9. We mention some studies, almost by chance:

   Susanne, C. (1977): Multifactorial inheritance and selection: analysis of the result of the test Progressive Matrices of Raven, Journal of *Hum. Evol.*, 6, 735-739.

   Dickens, W. T. – Flynn, J. (2001) Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychol. Rev.,* 108, 346-369.

   See also:

   Martinez, M. (2000): *Education as the cultivation of intelligence.* Mahwah, New Yersey, Lawrence Erlbaum Associates.

10. Csapó Benő (1987): *Representing the qualitative characteristics of reasoning by qualitative data two examples from the field of operational abilities*: *combinative and logical operations.* Bremen, Univ. Bremen.

11. Vári, P. (ed.): *PISA-vizsgálat 2000* (The PISA-2000 study). Műszaki Könyvkiadó, Budapest, 2003, pp. 146.