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 Empiricist* 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.

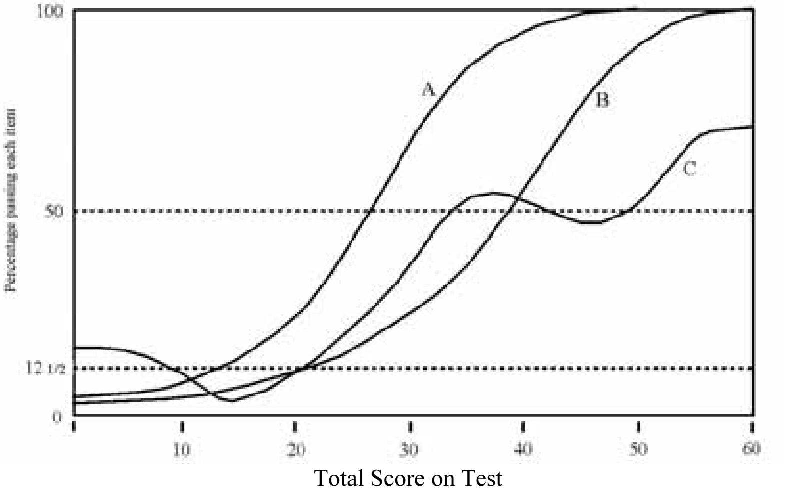
In the paragraph to which this note refers, "classical test theory" refers to the methodology most widely employed to construct what are deemed to be unidimensional tests. This usually involves intercorrelating the items and subjecting the resulting correlation matrix to some form of factor analysis.

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.

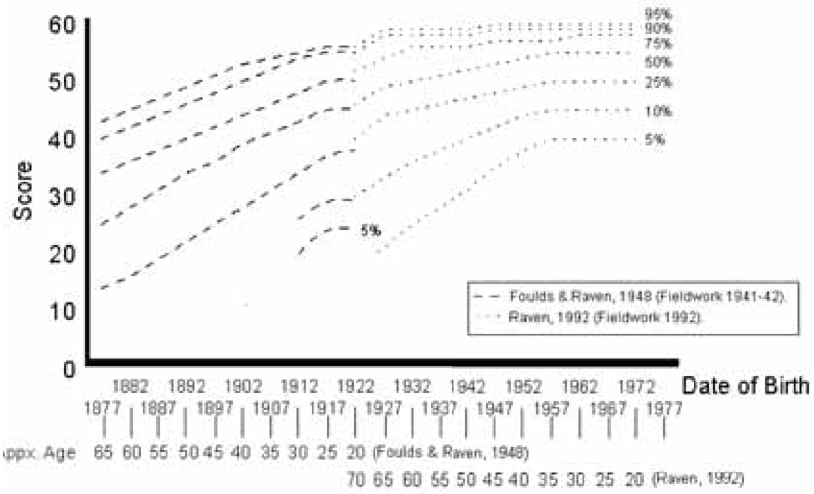
When data from a more difficult version of test – the *Advanced Progressive Matrices* (APM) – are examined it is, as shown in Table 7.1, immediately obvious that the scores attained by the more able *have* also increased dramatically. Unfortunately, we still do not know whether the effect is greater 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].)

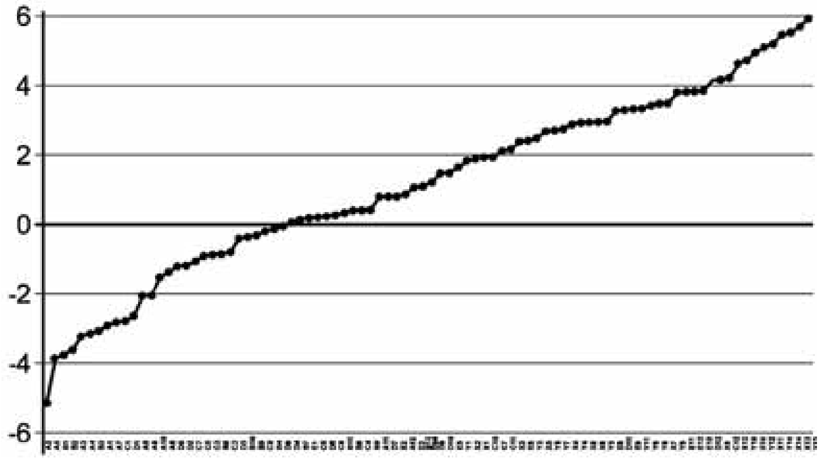
3. Problems arising from the available tests not yielding equal-interval scales

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.

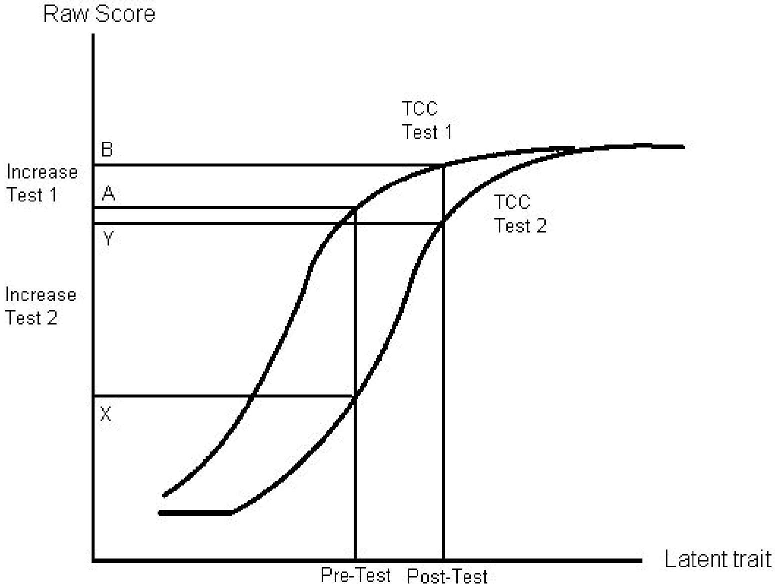
Our next step will be to elaborate on our earlier observation that, very surprisingly, the problem cannot be solved by developing a more difficult test, even a test conforming to the Rasch model.

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.)

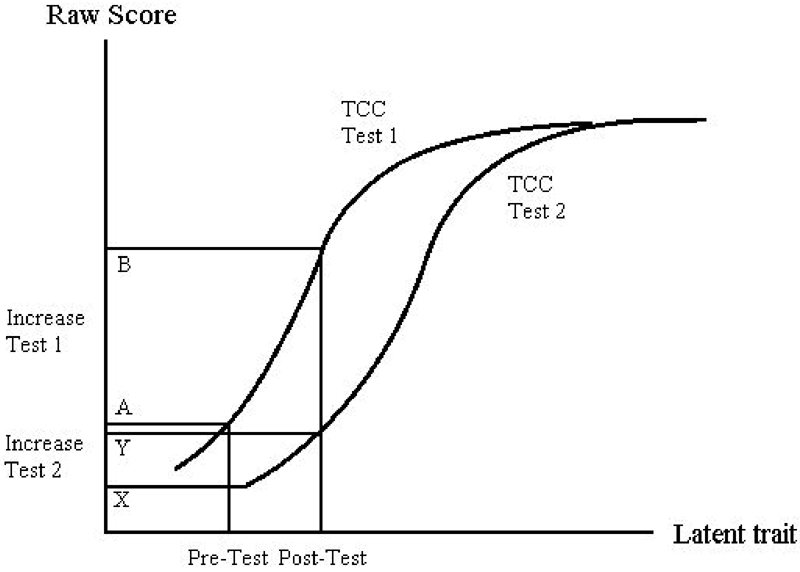
the low ability group are huge while those of the high ability group are trivial. On the other hand, if the researcher employs Test 2, exactly the opposite findings emerge.

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.

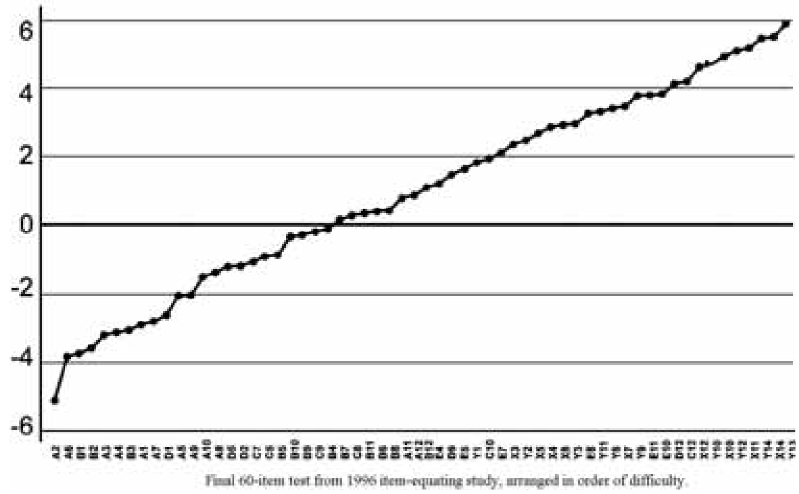
On the face of it, this provides a solution to our problem – provided such a test were used at both pretest and posttest. Unfortunately, not only does conformity to the Rasch model not ensure such an equal-interval scale, there can be no guarantee that, just because the overall distribution is as illustrated, 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.

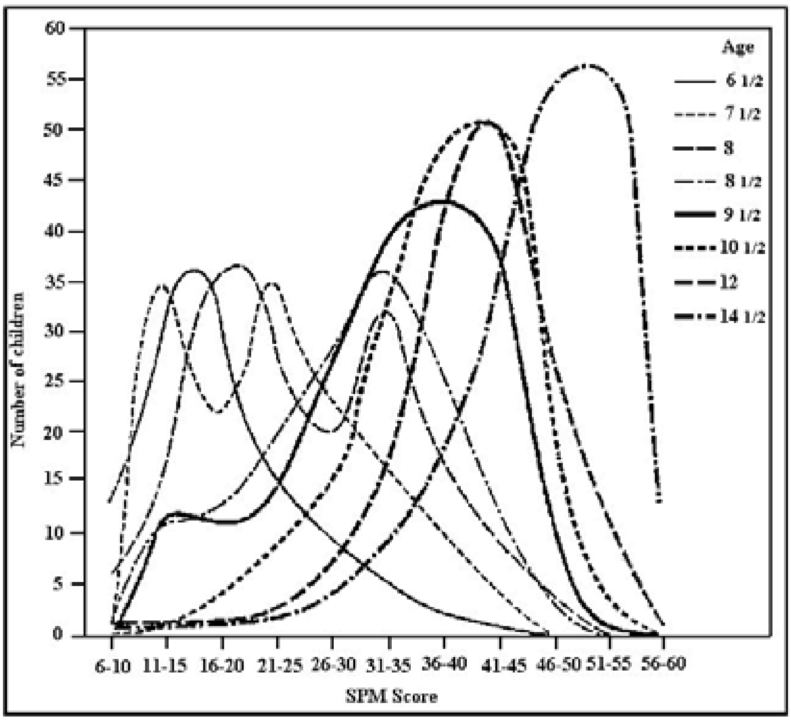
What if they were cumulated. Would we not then obtain a Gaussian distribution? The cumulated distribution is shown in Figure 7.8.

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.

******

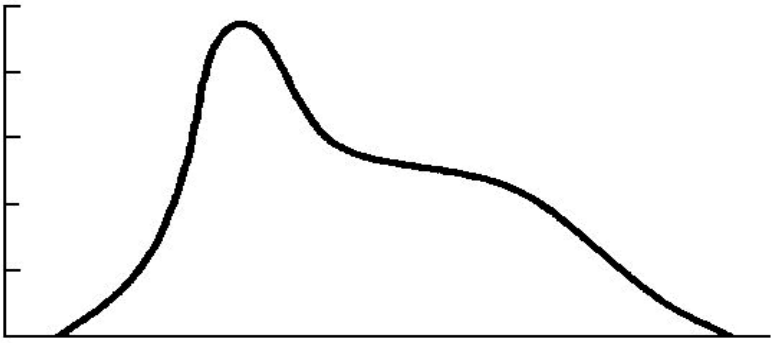
At this point it is instructive to return to Teasdale and Owen's observations. Although, in their text, they draw attention only to the fact that the scores of high ability respondents had increased hardly at all, it

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:

$$P(X_{ij}=1/\theta_j, b_i) = \exp\ (\theta_{vi}\text{-}b_i)/1 + \exp\ (\theta_{vi}\text{-}b_i)$$

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:

$$P(X_{\mathrm{vij}} = 1/\ \theta_{vi}, \beta_{i0}, \ldots, \overset{\mathrm{mi}}{\beta}_{imi}) = \exp(j\theta_v + \beta_{ij}) \ /\ \Sigma_{\mathrm{l=0}} \exp(l\theta_v + \beta_{il})$$

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 $0,\ldots,m_i$);

θ_v is testee *S_v's* person parameter;

β_{ij}, for j = $0,\ldots.,m_i$, 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.