Manual for Raven's Progressive Matrices and Vocabulary Scales

By

J Raven, J C Raven and J H Court

Section 3

Standard Progressive Matrices

(including the Parallel and *Plus* versions)

2000 Edition

With norms for the SPM *Plus* and formulae for calculating change scores

DEVELOPMENT OF PARALLEL & PLUS VERSIONS

Important Note:
This material forms part of the above manual and should only be used in conjunction with its General Introductory Section (General Overview).

Development of the Parallel and *Plus* Versions of the Tests

Background

Although normal levels of familiarity with the test have little effect on scores and although it is, in practice, easy to detect those who have been specially coached or who have memorised the answers[36], the view that the tests were becoming "too well known" had, by the mid 1970s, become sufficiently widespread to need to be taken seriously. Work was therefore put in hand to develop parallel versions of the tests. For reasons explained in Appendix 2, these initial attempts came to nothing. But, in the meantime, first the 1979 standardisation among young people and, later, the 1992 standardisation among adults in Great Britain revealed that it was necessary to add many more difficult items to the test to restore the discriminative power which it had had among more able adolescents and young adults in 1938. The way in which these parallel and extended items were developed is described in some detail in Appendix 2.

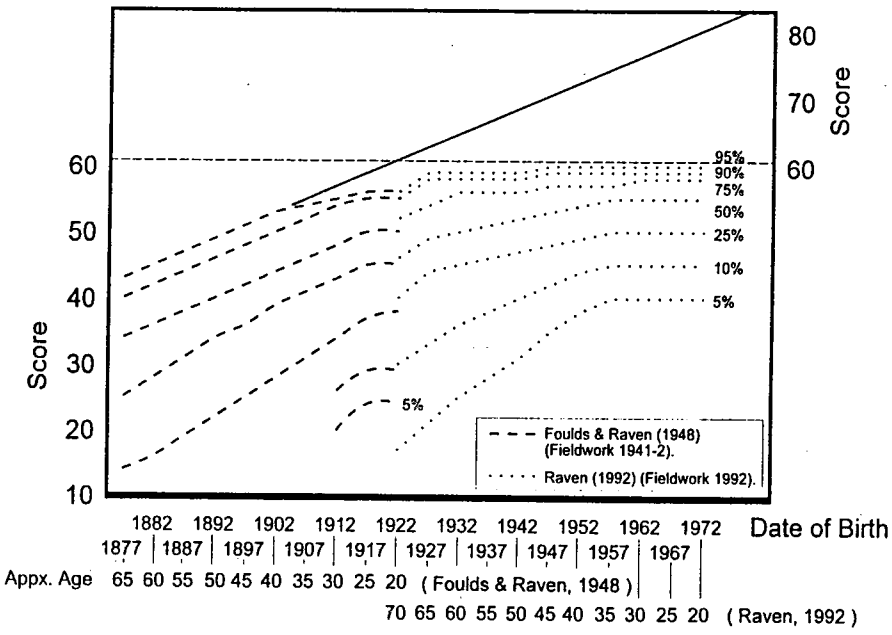
Here it is important to note that it was decided to try to develop a parallel test in which the items would match the old, item-by-item, both in overt solution strategy and in empirical difficulty. Only such a test would enable users to refer to existing normative data with confidence and ensure that any new data they collected could form part of the international data pool which has proved so invaluable in tracing changes in test scores over time and between cultures (and in this way documenting and explaining the impact of the environment).

However, the 1992 adult standardisation in Great Britain also had serious implications for the design of a version of the test with enhanced discriminative power at the upper end. These can be illustrated using Figure SPM1.

Figure SPM1 shows the mean and range of scores of adults born in each year from 1877 to 1972. It is apparent that there is a marked ceiling effect among today's young adults.

Extrapolating the almost-linear trajectory of the 95th percentile by date of birth from the point at which it begins to plateau (ie among those born in 1902) to 1974 reveals that one would need an 84-item test to achieve the same discriminative power among those of higher ability born in 1974 as the 1938 version had among those born before 1902.

Figure SPM1
Standard Progressive Matrices
Implications of Score Increase for Revised Test Difficulty
(Base Graphs Reproduced from Graph G2 in General Section)



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 1938 version.

In short, a test consisting of about 90 items would be required to restore to the SPM the discriminative power it had among more able respondents in 1938.

As described in Appendix 2, 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 was required for the item-equating study.

Results

The "Parallel" Test

Figure SPM2 graphs the raw-score distributions for the same abilities (expressed in logits) of the Classic and Parallel SPM tests. There can be no doubt that, at the level of total scores, the two tests are interchangeable.

Figure SPM3 plots the difficulty levels in logits of the old and parallel items of the SPM on a common scale. It is clear that, with the possible exception of

Figure SPM2
1996 Item-Equating Study
Ability (in Logits) Indicated by each Raw Score on Classic and Parallel SPM

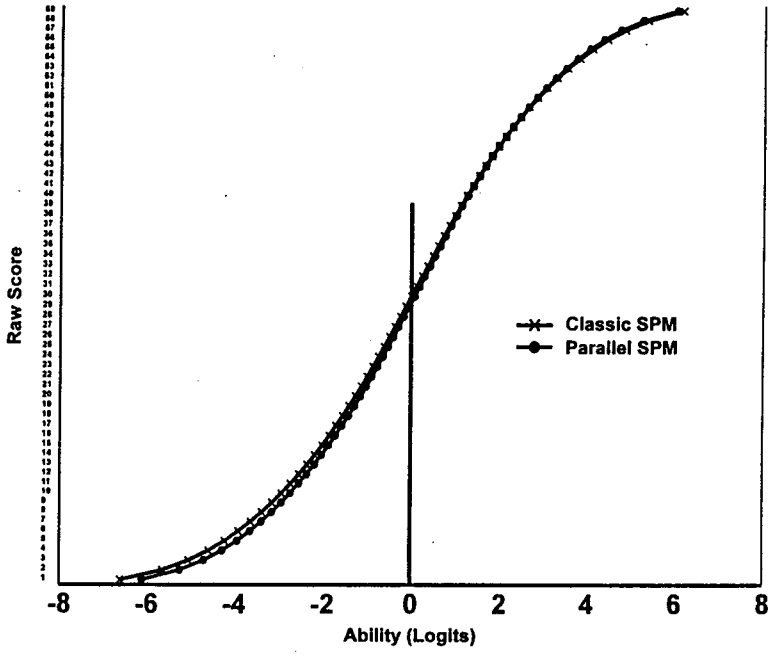
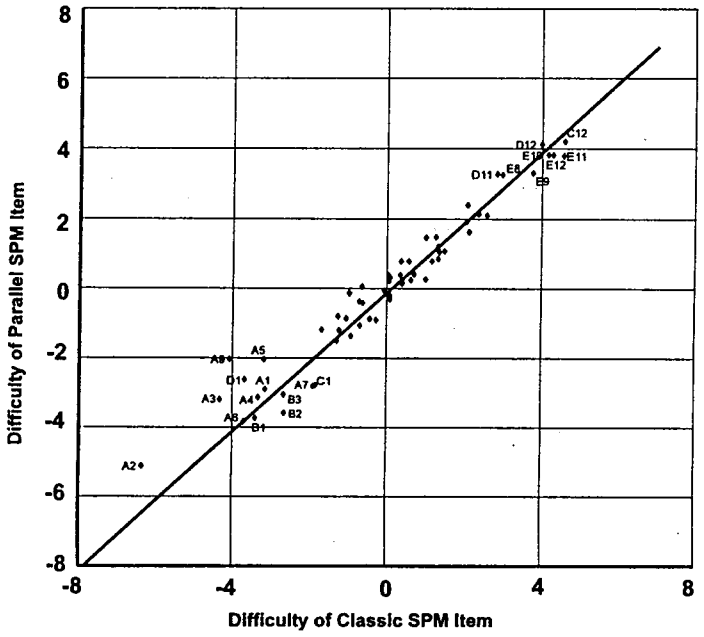


Figure SPM3
1996 Item-Equating Study
Comparative Difficulty (in Logits) of Classic and Parallel SPM Items



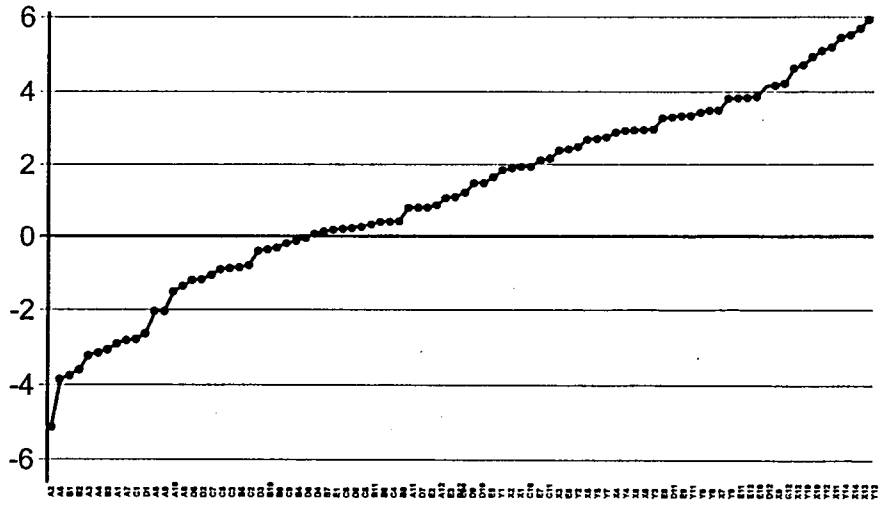
A9, the difficulty levels of the parallel items closely match those they replace. Inspection of the parallel A9 revealed the reasons for the mismatch and the item has now been modified.

SPM Plus

Turning now to the "extension" of the test to enhance its discriminative power and range of operational utility at the upper end, Appendix 2 describes the work which went into the development and trialling of a pool of new items and their reduction to 88 items for inclusion in the final item-equating study. Figure SPM4 shows the item difficulties of the 84 parallel and new items which remained after elimination of the four which had the poorest fit to the Rasch model.

Although it is not immediately obvious from a graph plotted on this scale, detailed inspection reveals that, in several sectors, there are a number of items having similar levels of difficulty. It followed that, by eliminating alternate items in these areas, a linear increase in item difficulty could be achieved. One of these areas comprises items D3 to A11. Clearly, by eliminating 24 items, largely from the set paralleling items from the original test, it would be possible to re-create a test having optimal length (in terms of fatigue and boredom) and yet discriminating across the entire range of intellectual ability. Such a test would have had another huge advantage. As Carver[37] has shown, a test constructed in such a way that equal increases in total score correspond to equal increases in the difficulty level of the most difficult items on which each score is based would

Figure SPM4
1996 Item-Equating Study
Item Difficulties (in logits): 84 Items – 60 Parallel Items and 24 Additional Items



have huge advantages. A test having a linear relationship between total score and *level of ability* indicated by the most difficult problem people were able to solve would prevent researchers drawing unjustifiable conclusions about such things as changes in the rate of maturation and decline of eductive ability with age. Apparent changes in *rates* of increase and decline of eductive ability with age stem from the fact that, as is apparent from Figure SPM4, the distribution of items by difficulty is uneven. The effect is that, at certain points in the distribution, 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.

Unfortunately, each of the Sets in the SPM (ie 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 recent Disabilities Acts.

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 SPM5 and, broken down by Set, in Figure SPM6. It is apparent from Figure SPM5 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.

This new version of the test, offering enhanced discrimination at the upper end, almost unchanged discrimination at the lower end, and a much more linear relationship between total score and ability, was named *SPM Plus*.

Figure SPM5
1996 Item-Equating Study
SPM *Plus*: 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

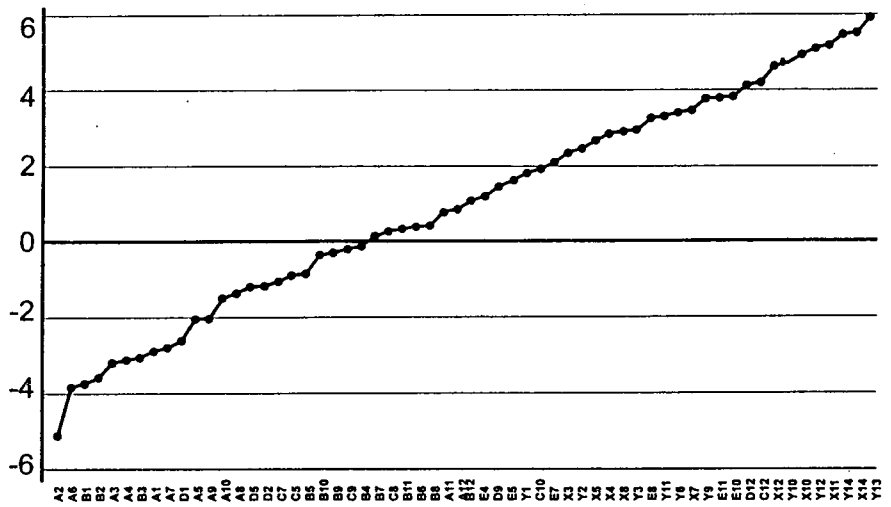
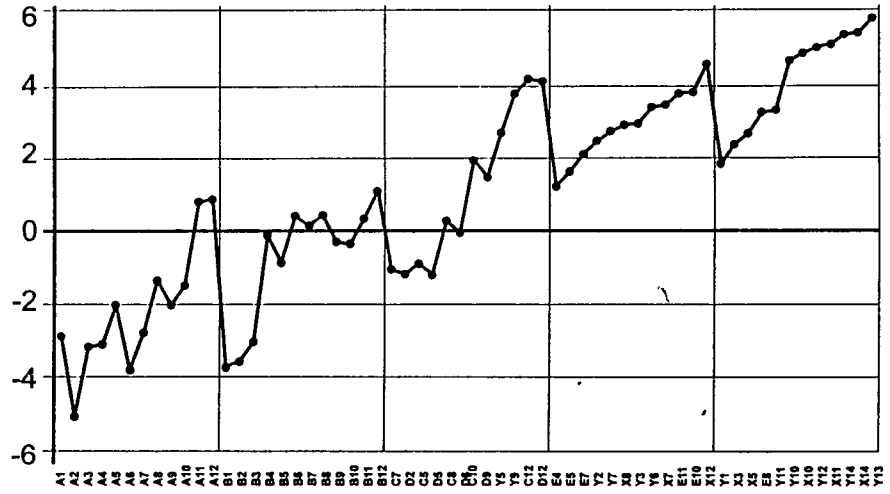


Figure SPM6
1996 Item-Equating Study
SPM *Plus*: Item Difficulties (in logits): 60 Items, Including ALL from Parallel Sets A and B and 5 each from Parallel Sets C and D, Arranged in Sets



Scoring Keys The final step taken to achieve the objective of preventing anyone who might have memorised the correct answers to the SPM capitalising on that advantage was to change the positions of the correct answers. *The Scoring Keys for both the Parallel version of the SPM and SPM Plus therefore differ from that for the Classic SPM and from each other.*

Ex Raven, 2000
SPM Manual.

Appendix 2: Development of the Parallel SPM and SPM *Plus*

Early attempts to develop parallel versions of the tests

As explained in the main text, the view that it would be desirable to introduce a parallel version of the test had become accepted by the mid-1970s. An attempt was therefore made to find researchers who could help us with the work.

Jacobs[278] had studied the mental processes employed to solve the problems, coming to the conclusion that these involved such things as "flip over", "rotation", "subtraction", and so on. More importantly, he had concluded that item difficulty depended on the number of these processes that had to be undertaken simultaneously. Jacobs was therefore asked to prepare a set of items to test his theories (and find out whether there was a basis for commissioning him to undertake further work). These items were included in the 1979 standardisation of the SPM among young people in Britain. It turned out that his items varied little in difficulty. This route to the development of parallel versions of the tests had therefore to be abandoned.

A largely unexpected result of the main 1979 standardisation was, however, the discovery that test scores had been increasing quite markedly over the years. This led to recognition that the real need was not so much for "parallel" versions of the tests as for versions which would restore their discriminative power among the more able.

Starting in 1984, the author had extensive correspondence with David Andrich at the University of Western Australia. Andrich was conducting a longitudinal study of cognitive development as assessed by the SPM and APM and a series of Piagetian tasks[279]. The study involved the practical deployment of item response theory and the development of computerised and adaptive versions of the tests. These exciting developments brought the author to Perth in September 1988.

In the course of this visit it emerged that:

a) Andrich's team had already developed a complete parallel test by making minor changes to the existing items. In the process, they had discovered that seemingly slight modifications often resulted in major changes in difficulty.

b) One component of the project involved talking to children as they solved the problems with a view to gaining insight into the strategies they used and then using this information to generate new, more theoretically-based, items.

c) It had proved unexpectedly difficult – indeed almost impossible – to develop items more difficult than E11 and E12 of the SPM or the more difficult items of the APM.

Discovery of a project covering so much of what was involved in developing new items led first to a discussion of exactly what was required and then to commissioning Andrich's colleague, Irene Styles, to carry out the work.

Evolution of project design and contributions of collaborators

It was agreed that the first thing to do was to develop a new test which would parallel the old on an item-by-item basis both in overt solution strategy and empirical item difficulty. This was important because it would mean that users would be able to refer to all existing normative data with confidence and that any studies carried out with the new test would yield data which would form part of the international data pool, which has proved so invaluable in tracing changes in test scores and norms with time and culture and, in this way, documenting and explaining the impact of the environment.

At least 12 *new* items were also required to restore the SPM's ability to discriminate among those with higher scores.

It was also proposed that work would later be put in hand to develop new items having very different surface properties but based on a clearer understanding of the psychological processes involved.

It was decided that Styles would run a series of pre-pilot studies on samples of about 80 children, but the very large studies required to check the equivalence of the old and new items would be arranged by the author. At the time, it was envisaged that these larger studies would also be conducted in Australia with the co-operation of the Australian Council for Educational Research , who had recently conducted the previously-mentioned Australian standardisation.

Owing to changes which were shortly thereafter introduced into the Australian universities, development of the parallel and new items took much longer than expected and the proposal to conduct the large-scale item-equating study in Australia fell through.

In the event, while the parallel items were acceptable, they were felt to be rather clinical in appearance and to lack the bold drawing, intrinsic interest, and artistic flare which had been deliberately introduced by J.C.Raven when developing the original items by employing an artist (Henry Collins) to help him meet the specification he had developed[280]. Steve Hughes, who had developed the Macintosh computerised version of the tests used in the Minnesota Twin Study, was therefore commissioned to rework them.

More seriously, as predicted from the previous experience of Andrich's team, the development of more difficult items turned out to be problematic. Worse still, by the time Styles' items became available, the results of the 1992 standardisation among adults had, as explained in the main text, revealed that many more difficult items were required than had originally been envisaged.

Given that Styles had been asked to create only 12 new items, how were these to be generated?

Identification of more difficult items to be paralleled

The challenge was taken up by J.C.Raven's grandson, Michael Raven, who proposed that the obvious answer was "by developing items parallel to those of appropriate difficulty selected from the APM". He bullied the author into thinking out how to identify such items, and then set about paralleling them himself[281].

How were the APM items which would need to be paralleled for use in a new SPM to be identified? Items having specific difficulties – not just any old items – were needed if both the linearity in the increase in scores with age that is evident in the left hand portion of Figure SPM1 and the separation between the raw scores corresponding to each percentile were to be maintained.

Fortunately, work which had also been carried out by Andrich and Styles in the interim[282] was available to help solve this problem.

Andrich and Styles had mapped the item difficulties expressed in Rasch logits of both the SPM and APM items onto a common scale (see Table RS4C2 in *Research Supplement No.4* and Table SPM5 in the Reference tables in this [SPM] Section).

The way in which this table was used to identify the APM items to be "paralleled" to enhance and extend the discriminative power of a new SPM is best illustrated by working though an example.

Reference to Figure SPM1 shows that the test's ability to discriminate among those aged 45 and under was, in 1992, for all percentiles above the 50th, much less than it had been in 1942. For example, the gap between the 50th and 75th percentiles for this age group at the time of the 1992 adult standardisation was about two items less than it was in 1942. The difference between the 75th and 90th was only one item where it had previously been about four. And so on.

It followed that new items were required all the way from the 50th percentile upward, and not only above the top end of the 1938 edition of the SPM.

Reference to the Figure indicates that the first item needed would be one which would increase the score needed to reach the 95th percentile for those aged 35 or younger in 1942.

It is obvious that, once such an item was included, it would also raise the score needed to attain the 50th percentile for those aged 40 and younger in 1992. And, of course, the scores required to reach all the higher percentiles.

Looking again at the Figure, it is evident that the required item would be one which would raise a score of 54 to 55. That is, it needed to be just a little easier than the most difficult item contributing to an SPM score of 54. Inclusion of such an item would also have the effect of moving the graph for the 90th percentile for 25 year olds in the 1942 standardisation one up. It would also move up the graphs for these percentiles for all age groups in 1992. Finally, it would move up the 75th percentile for everyone under 60 in 1992.

In actual practice, this exercise was carried out using the tables on which Figure SPM1 was based rather than the graph itself.

The 1942 data which contributed the graphs on the left of Figure SPM1 come from Table SPM V in the 1988 edition of this (SPM) Section to this *Manual*. The 1992 data come from Table SPM7 in this edition of this Section.

Table SPM V in the 1988 edition shows that the difference between the 90th and 95th percentile falls from three to two items at age 35. It follows that what was needed was one more item whose difficulty corresponded to the most difficult item contributing to total scores between 51 and 53.

Reference to Table SPM2 shows that a score of 52 is typically achieved by a combination of scores involving 10 from Set C, 10 from Set D, and 8 from Set E. So, on the assumption that the items are arranged in the correct order of difficulty, it follows that we needed an item of equivalent difficulty to C10, D10, or E8.

Table RS4C3 (in *Research Supplement No.4*) shows that the APM item closest in difficulty to SPM C10 is APM 5, SPM D10 (which seems to be easier than C10), corresponds to APM 9 (which, by the same token, seems to be easier than APM 5), and SPM E8 corresponds to APM 13. The particular item which was chosen then depended on the context into which it would fit.

Having selected an item which, if included, would have produced the desired result, all scores of 52 and above in the norm tables were increased by one. Thereafter, the next convergence between the lines of figures was identified and an item which would correct it sought in the same way.

The whole process was then repeated until enough APM items had been identified which, if included in the SPM, would have resulted in a linear increase in the percentile norms with date of birth, while maintaining the separation between the curves.

The logic behind each of these APM items was then examined and used to generate a set of parallel items which would be expected, if the logic had been correctly identified, to have similar difficulty levels.

The next problem was to identify a number of still more difficult items which would provide the scope for increases in scores over time that the 1938 edition of the test had in 1942.

This was more problematic than might have been expected.

Reference to Graph APM7 (in the *APM* Section of this *Manual*) shows that even the APM, by 1992, had only three items which were too difficult for 95 percent of the population.

To restore the original properties of the SPM, about five such difficult items were needed.

As has been mentioned, Styles had, even before the author made contact in Australia, discovered that it was extremely difficult to develop items more difficult than those in the existing SPM and APM.

By the time Michael Raven came up against this problem, Linda Vodegel-Matzen had already been contracted to develop such items for an extended *Advanced* Progressive Matrices. Before this, she had

developed a parallel version of the SPM[283] which "worked" in the twin sense of having a good range of item difficulties and a high correlation with the SPM by applying the rules of Carpenter and Just[284]. Unfortunately, although her test had a number of advantages – including a Carpenter and Just logical basis for *all* items and theoretically based construction of distracters – it did not correspond to the original at an item by item level and therefore could not become the test we needed.

The preliminary items she had developed for the APM were therefore reviewed to identify a number which might be included in the new *Standard Progressive Matrices*.

Trialling the items

As a result of all this work we now had:

- A set of 12 items overlapping in difficulty with, but going beyond, Set E of the SPM that had been developed and trialled by Styles.
- A set of items of apparently equivalent logic and thus, it was hoped, equivalent difficulty to existing APM items which, had they been included in the SPM, would have restored to it some of the properties it had in 1938.
- A number of items developed by Vodegel-Matzen which were expected to be more difficult than the existing APM items.

It was arranged to administer all these new items along with a number of marker items from the existing SPM and APM to groups of "semi-volunteers" from the University of Amsterdam. To keep the time the students were required to devote to the project to a reasonable level, two sets of 23 items were developed and each administered to 35 students. Set I of the existing APM was used as a familiarisation and practice exercise.

Consideration of the data from these trials resulted in the selection of 28 new items which were potential candidates for inclusion in the new test. These were divided into two sets of 14 for the large scale studies which followed.

It was agreed that Styles would use her Rasch scaling programmes to map the item difficulties of the parallel items, together with those of the new items which had been developed to enhance the discriminative power and range of applicability of the test at the top end, onto a common scale based on the item difficulties of the originals – all expressed in logits.

Both Andrich and Hambleton[285] had, by this time, demonstrated that Rasch parameters become very unstable when derived from samples of less than 600. Clearly it was going to be necessary to test large numbers of respondents. Despite these conclusions of her colleagues, however, Styles was satisfied that it would be possible to make do with smaller numbers *provided the same people were given each old and new item* because this would make direct comparisons possible. It would then not be necessary to depend on having numbers large enough to minimise errors arising from sampling variation.

To give an idea of the numbers which would otherwise have been required, it may be mentioned that the difficulty levels presented in Table RS4C2/SPM5 were calculated from data supplied by approximately 600 MENSA applicants who had taken both the SPM and the APM, a random sample of a 3,000 children of all ages who had taken the SPM, and another 1,000 who had taken only the APM.

Unfortunately, administering items having exactly the same logic, but presented in a different form, to the same children seemed likely to introduce both practice and fatigue effects. An attempt was therefore made to counteract these by administering the items in different combinations. It was arranged to administer different subsets of items to 5–8 year olds, 11–13 year olds, and 16–18 year olds in such a way that it would be possible for subsets of each of these groups to complete the alternate tests a month or so later.

It is important to note that, because this was an *item-equating* study, not a norming study, it was not necessary to test a random sample of children and young people of all ages. On the contrary, 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.

We may once more take an example – this time a hypothetical one – to highlight the implications of this. 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 as the ICCs.

Percentages calculated on bases of ten are clearly extremely unreliable.

As if this problem were not bad enough, if a *random* sample of young people were tested, the numbers obtaining each total score from 1 to 60 would be anything but equal. Many would get scores around the average and there would be very few obtaining scores in the tails of the distribution. Consequently, the bases for the percentages of these low and high scores 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, over-represented. Put another way, an ideal distribution for our work would be 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 SPM28.

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 smaller subsets 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 the table which follows, 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.

Before looking at the Table, one more word of explanation is necessary. Although it has not been necessary to mention it before, the study included the *Coloured* Progressive Matrices as well as the Standard version since Sets A and B are common to both tests.

After a number of abortive attempts, it was arranged that the testing would be carried out in the Netherlands. Unfortunately, Vodegel-Matzen, having obtained her PhD, left the University. The testing programme was then taken over by Rieneke Visser and Saskia Plum. Because the 18 year olds required by the design were, at that time, involved in examinations, it proved difficult to obtain their co-operation. Fortunately, Francis Van Dam and Dr. J.J.Deltour were able to arrange for large numbers of students of the right age to be tested in Brussels and Liege. In addition, because it looked for a while as if the project in the Netherlands would grind to a halt, arrangements were made for Anita Zentai to test children in Hungary. In the event, the Dutch team also filled their quota by the time the Hungarian data came in and, as will be seen from Table SPM28, we ended up with considerably larger numbers than had been anticipated.

Table SPM28
Sample Design for 1995 Item-Equating Study

Booklet number	Target age	Sets covered	Arrangement of items	Total no. of items	Target no. of respondants	Target to retest	Actual no. tested
Coloured Progressive Matrices							
1	5½–8½	A Ab B	OA1 NA2 OA3 NA4 ... OAb1 NAb2 ... OB11 NB12	36	150	25	287
2	"	"	NA1 OA2 NA3 OA4 ... NAb1 OAb2 ... NB11 OB12	36	150	25	274
3	"	"	NA1–NB12	36	150	25	373
3B	"	"	OA1–OB12	36	150	25	164
Standard Progressive Matrices							
4	11 12 13	A–E	OA1 NA2 OA3 NA4 ... OE1 NE2 ... OE11 NE12	60	150	25	238
5	"	"	NA1 OA2 NA3 OA4 ... NE1 OE2 ... NE11 OE12	60	150	25	240
6	"	"	NA1–NE12	60	150	25	224
6B	"	"	OA1–OE12	60	150	25	215
Restoration of discriminative power of Standard Progressive Matrices							
7	16–19	D E X	ND1–NE12 + 1st 14 new items	38	300		343
8	"	D E Y	ND1–NE12 + 2nd 14 new items	38	300		267

Results of the item-equating study

The main results from the item-equating study have been summarised in the section entitled "Development of Parallel and Extended Versions of the Tests". However, the decision to publish one "parallel" test and a 60-item test containing more difficult items was there presented as a more or less obvious conclusion. In fact it was a decision which involved considerable agonising.

Having achieved our basic objective of developing an 84 item test which would "restore" the discrimination at the upper end which the original test had among those tested in 1938, the question was: Should we publish:

1. Only one new test consisting of 84 items which would both exactly parallel the old test and, at the same time, extend its discriminative power.
2. Two new tests, one exactly paralleling the old, and the other, starting with the new Set C, but including two new Sets – tentatively called Sets F and G.
3. Only a new "condensed but extended" test of 60 items retaining the structure of Sets A and B (and, to some extent, C and D) but including many more difficult items. Scores on such a test would only be convertible to those on the original test through conversion tables.
4. A directly parallel test and a "condensed but extended" test of 60 items (ie tests 2a and 3 above).

Theoretically, one more option was open to us. This was to publish a single test eliminating the "easier" items at the bottom end and thus restoring to the test the properties it had among those tested

in 1938. This option was rejected because the data collected in the 1992–93 adult standardisations clearly show that the test is now working well among older adults where – as can be seen from Figure SPM1 - the lower percentiles from the 1942 study became meaningless. Furthermore, this lower-scoring group had, as a result of changes in the social context of testing and, more specifically, recent Disabilities Acts, become of major interest. In addition, both the current edition of the test and its parallel were now working well among young children. The tests limitations were confined to the top end of the adolescent and young adult population.

After canvassing the opinions of those most involved with the test, and influenced by the work of Carver[286], it was decided to publish both an exactly parallel test and a "condensed but extended" test retaining all the items paralleling those in Sets A and B, five each from C and D, and the remainder drawn from Set E plus F and G from the item-equating study.

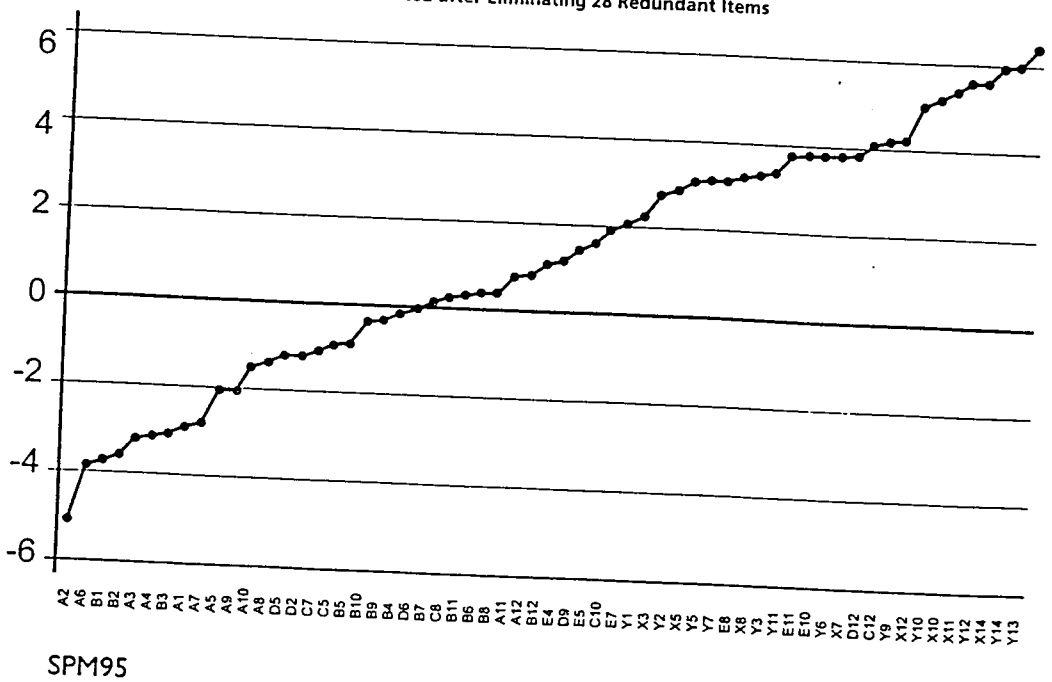
A coda: instability in Rasch parameters

It is well known that Rasch set out to develop item metrics which would be independent of the other items included in the item pool and the ability of the population tested. It is widely believed that this problem has been largely solved (albeit that one needs large numbers to achieve stability).

In point of fact, in the course of the present study, we encountered *huge*, and still largely unexplained, variance in item parameters depending on which of the data sets identified in Table SPM28 were included in the analysis. The variation was so great that it proved necessary to, for example, lock the item statistics for Sets C, D, E, X, and Y when entering or removing the item-person data sets for Sets A and B.

Perhaps one of the most perplexing findings was that the parallel items *in Sets A and B* whose difficulty indices did not appear to match the Classic items they were designed to parallel – ie which did not fall on the regression line in the CPM equivalent of Figure SPM3 – varied with whether or not the items *for Set Ab* were included in the analysis. This in spite of the fact that the population of respondents on whom the statistics were based was identical.

Figure SPM7
1996 Item-Equating Study
SPM *Plus*: Item Difficulties Re-calculated after Eliminating 28 Redundant Items



The point may be illustrated here using the *least* troublesome of these difficulties. Figures SPM5 and SPM6, although they relate to the 60 items finally retained, are based on item statistics derived from the analysis of the 88 items carried forward into the final item-equating study and for 84 of which statistics are shown in Figure SPM4. When item statistics for the same items from Sets C, D, and E only (the item parameters for Sets A and B having been locked for the reason already given) were re-calculated (using the same data from the same population of respondents) after eliminating the 28 apparently redundant items, Figure SPM7 was obtained. Note that not only is the approximation to a straight line not so close, the *position* of some items in the apparent rank order of difficulty of the items has changed.

It remains to note that our experiences in this analysis not only seem to have major implications for those who rely on Rasch analyses, they also have implications for the weight to be attached to the studies based on different populations (of different ages and ability levels) which have claimed to indicate a need to re-order the items in the Classic SPM.