## Chapter 3

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

John Raven

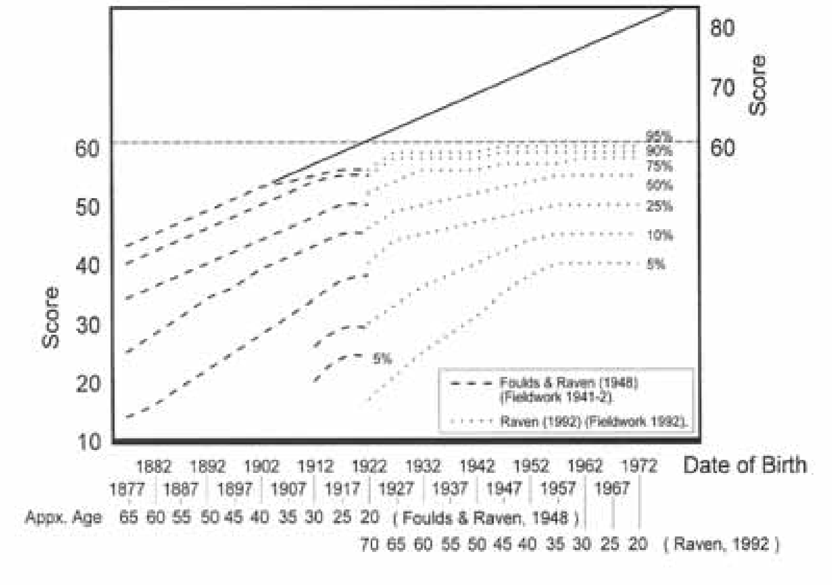
As we have seen, the development of the *Standard Progressive Matrices* ***Plus*** (SPM+) was precipitated by the dramatic and unexpected international increase in RPM scores that had taken place over the years. This resulted in the failure of the *Classic* Standard Progressive Matrices to discriminate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

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

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

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

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

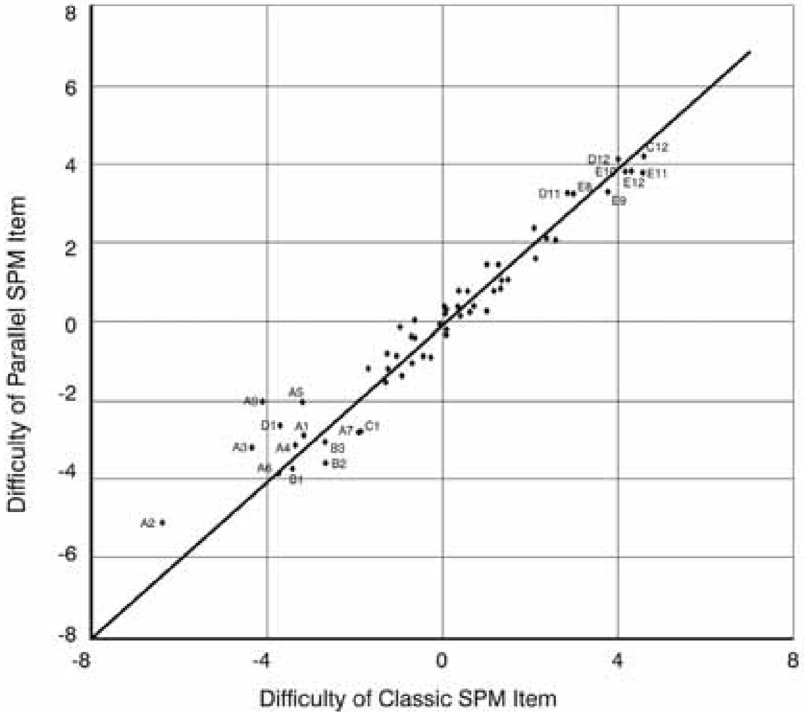


As described in Appendix 2 to the 1998-2004 editions of the SPM Section of the *Manual*[3.1], the energies of numerous people in several countries were harnessed to the task of developing the required items, conducting and analysing pilot studies, and finally testing the large number of people at all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

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

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

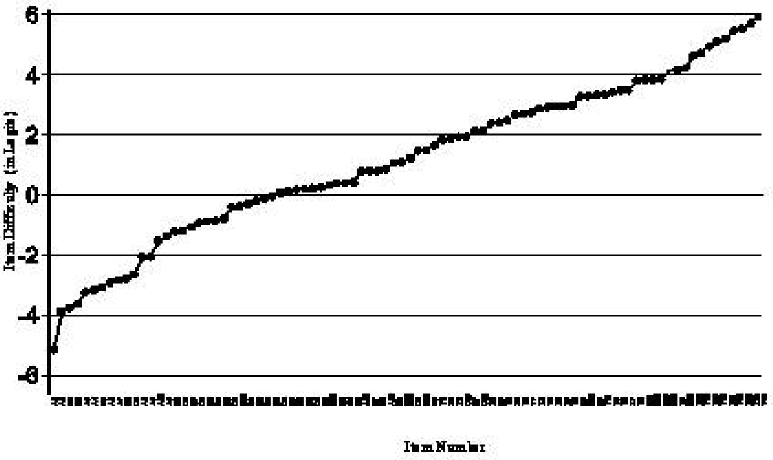
Figure 3.2. **Comparative Difficulties of *Classic* and *Parallel* Standard Progressive Matrices Items**
(Based on 1996 Item-Equating Study)



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

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

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



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

Unfortunately, eliminating items to leave only those that result in equal increments in difficulty poses problems because each of the Sets in the *Classic* and *Parallel* versions of the SPM (i.e. A, B, C, D, and E) is made up of items of a different type. These not only require different forms of reasoning but also introduce those being tested to the logic required to solve the next most difficult item in that Set. Elimination of the clearest candidates for removal would have resulted in a selection of 60 items which would have destroyed this unique property of the test. It would also have destroyed the comparability between the SPM and CPM. And it would have reduced the test's new-found ability to discriminate well among older adults and young children in zones where the 1938 version of the test did not work too well.

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

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

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

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

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

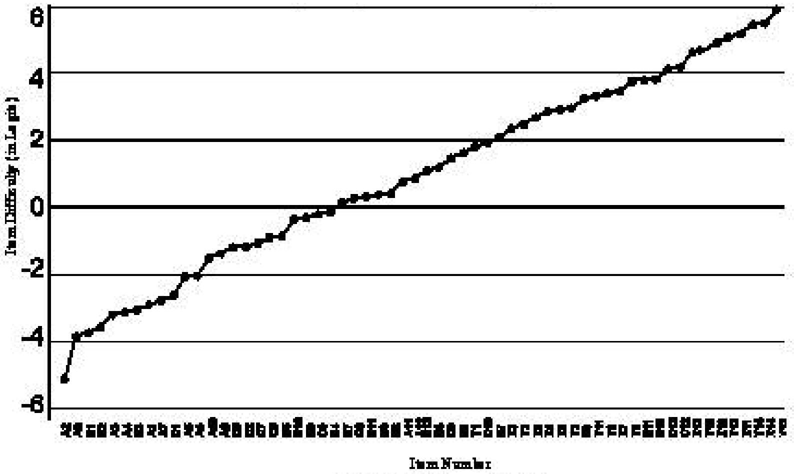
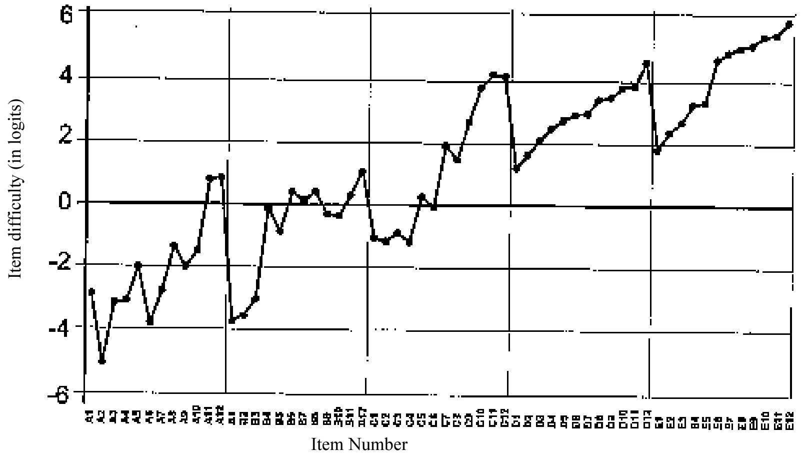


Figure3.5. *Standard Progressive Matrices Plus* 1996 Item-Equating Study
**Item Difficulties (in Logits)**
**60 Items, Including ALL from Parallel Sets A and B and 5 Each from Parallel Sets C and D, Arranged in Sets**



# Appendix

### *The Design of Samples in Test Development*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

**Table 3.1. Sample Design for 1995 Item-Equating Study**

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

## Notes

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

3.2. Carver (1989)

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

## References

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

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