USES AND ABUSES OF INTELLIGENCE

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

John and Jean Raven
(editors)

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

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

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

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

PART IV Outstanding Conceptual and Measurement Issues

PART V: Emerging Applications

PART VI: Some Outstanding Ethical Issues

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

Chapter 5

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

John Raven, Joerg Prieler, and Michael Benesch

Abstract

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

******

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

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

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

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

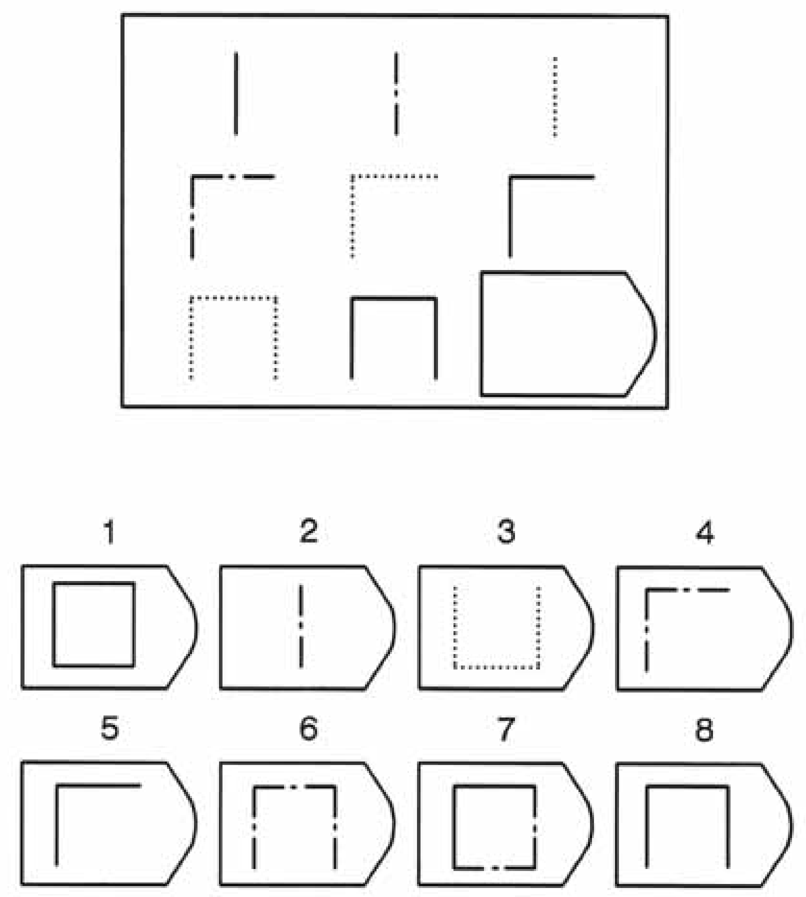
Background

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

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

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

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



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

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

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

Unfortunately, as shown in particular by Flynn (1984a&b, 1987), Raven (1981, 2000b), Raven, Raven, and Court (1998, updated 2003), the scores achieved by samples of the general populations of many countries on the Raven Progressive Matrices (RPM) tests have been increasing dramatically over the years. 50% of our grandparents would be assigned to Special Education classes in the US if they were evaluated against today's norms. [As an aside it is important to note that this increase has been documented on many measures of eductive (but not reproductive) ability, whether verbal or non verbal, for many other abilities (such as athletic ability), and for many other human characteristics such as height and life-expectancy. Readers interested in reviewing the evidence showing that these increases are not due to any of the obvious causes may find Raven, Raven, & Court (1998, updated 2004) and Raven (2000a&b) of interest.]

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

The Measurement Model

Although it is well known that the items in the RPM tests become progressively more difficult (albeit in a cyclical format [which was introduced to provide training in the method of working]), it is not widely known that the *Standard Progressive Matrices* (SPM) was initially developed using a graphical version of what has since become known as "Item Response Theory". For example, in an article published in 1939, J. C. Raven included sets of graphs of the form that have since become known as Item Characteristic Curves (ICCs) for both the *Coloured* and *Standard* Progressive Matrices tests. These have been reproduced in the Introductory Chapter to this book. Similar graphs for the *Advanced* Progressive Matrices were included in the *Guide* to the use of that test which was published shortly after the Second World War (Raven, J. C., 1950). The Graphs in these articles (which correspond to those in Figure 5.4 below) plotted, for each item, the percentage of respondents with each total score who got the item right. The graphs for all items in the overall test (or the sub-set under investigation) were, as in Figure 5.4, included in a single plot so that they could be examined for cross-overs, spacing, and coverage of the domain of ability it was hoped to assess. The objective was to select items whose curves had smooth ogives (instead of wandering all over the place), had ogives of approximately the same form, were equally spaced, and probed the whole domain of ability for which the test was intended. J.C. Raven argued that wandering ogives indicated that the items concerned were faulty. For example, there might be some feature of the item which confused more able respondents and distracted them from the correct answer. In a perfect world, the ogives would be vertical and equally spaced. One would then have the level of measurement achieved in a meter stick or foot-rule. There would be a 1 to 1 relationship between total score and final item passed.

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

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

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

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

At this point we may draw attention to the way in which we have been using the term "Item Characteristic Curve". We are aware that some measurement theorists would like to restrict the term to graphs produced *after* transforming the data applying some mathematical variant of IRT (and, more specifically, plotting score on the *latent* variable being measured by the test, instead of raw score, on the horizontal axis[5.2]). However, as we shall shortly show, such graphs typically render crucially important information invisible. 1-parameter models, for example, conceal what is happening to the proportions getting the item right before the item begins to be problematical to a significant proportion of those tested and differences in the slopes - discriminative power - item-total score correlations - of the items. To avoid confusion we have, in the remainder of this article, referred to the kind of graphs that Raven produced as “empirical” ICCs.

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

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

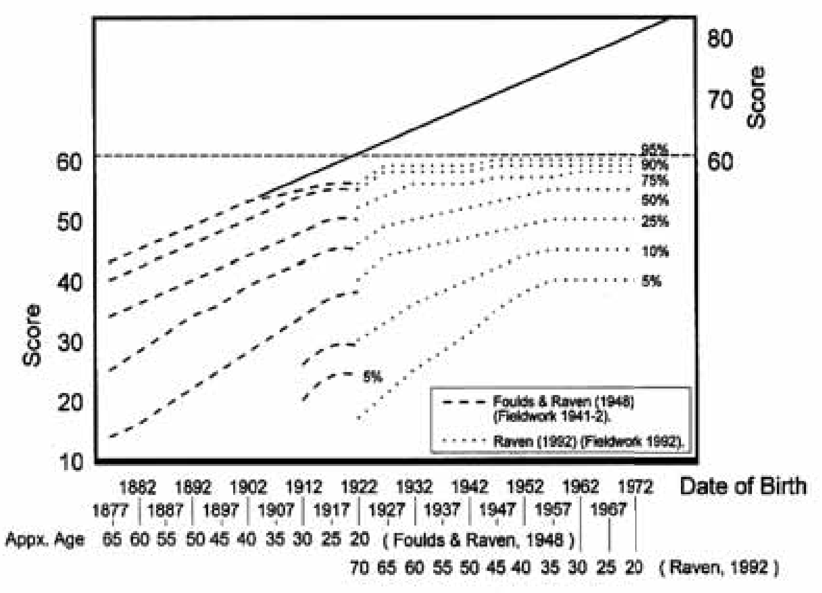
The Development of the SPM *Plus*

It turned out that the development of more difficult items for the SPM was no easy matter. Despite Vodegel-Matzen’s (1994) outstanding work, it gradually became clear that there was much more to Raven’s items than met the eye, and certainly a great deal more than Carpenter, Just, and Shell (1990) would have us believe. Indeed items generated for us by a widely cited authority on the rules governing the understanding and solution or Matrices items (who we will not name here) did not scale at all! The assistance of Irene Styles, Linda Vodegel-Matzen, and Michael Raven was therefore recruited. At first it was thought that the addition of twelve more difficult items would be sufficient to restore the discriminative power that the SPM had had among more able respondents when it was first published. However, it gradually became clear that twice that number were required (see Figure 5.2, in which the graph showing the increase in the 95th percentile from those born in 1887 to 1912 has been extrapolated to 1982). Since we did not wish to modify the original SPM (for which such a vast pool of research data from so many countries existed), we also set about paralleling the existing items and checking that the proposed parallel items not only had equivalent difficulty to the old ones but also worked in the same way. To achieve these ends, a series of pilot studies of different sub-sets of old and new and more difficult items were undertaken. These were mostly conducted on about 300 respondents whose ages and ability levels seemed appropriate from the point of view of trialling the items concerned. The data from these studies were then analysed by Styles using 1-parameter mathematically-based IRT programs and the results used to whittle down the total pool of items to 108 that were carried forward into a full-scale item analysis. (The process is described in greater detail in Raven, Raven & Court, 2000/04.)

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

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

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



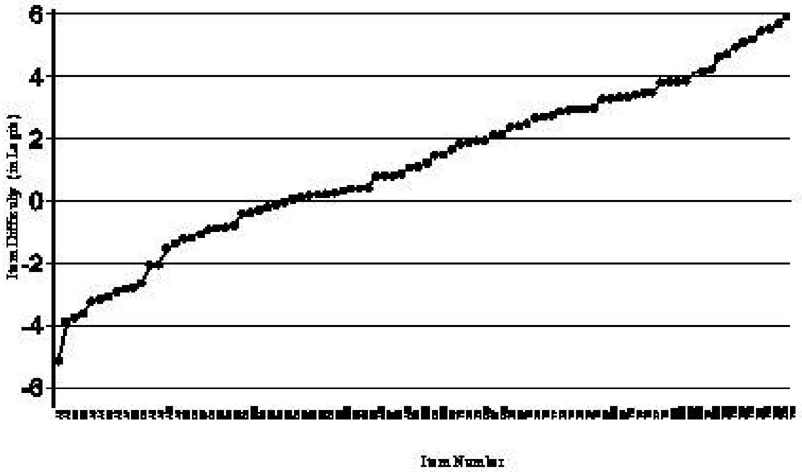
booklets from different samples of respondents in the analysis.

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

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

We used the item-statistics she sent us to first reduce the total number of items from 108 to the 84 we thought we needed for the new test. However, as can be seen from Figure 5.3, a graph of the item difficulties of those 84 items revealed a number of plateaux (e.g. between items D3 and A11) where there were several items of similar difficulty. It was obvious that, if some of these items could be eliminated, we could re-create a 60 item test in which the ability of the *Classic* SPM to discriminate at the upper end would have been restored without destroying its new-found ability to discriminate among the less able.

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



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

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

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

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

The Romanian Study

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

Particularly because it covered such a wide range of ability, this study provided us with a superb opportunity to replicate and extend the item analysis that had been carried out whilst we were developing the SPM+ test. This was particularly important because, in that study, data relating to the items finally retained were collected when those items were presented to respondents in the context of different sub-sets of items, many of them of similar logic and difficulty drawn from the *Classic* SPM. Respondents' answers to the new items could well have been influenced by this context. The size and coverage of the Romanian sample not only goes a long way toward counteracting some of the problems known to be associated with calculating item statistics for the easier and more difficult items, it also meant that despite the, inherently unstable, nature of IRT-based item statistics (Hambleton et al., 1991) there was a reasonable chance of obtaining meaningful data.

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

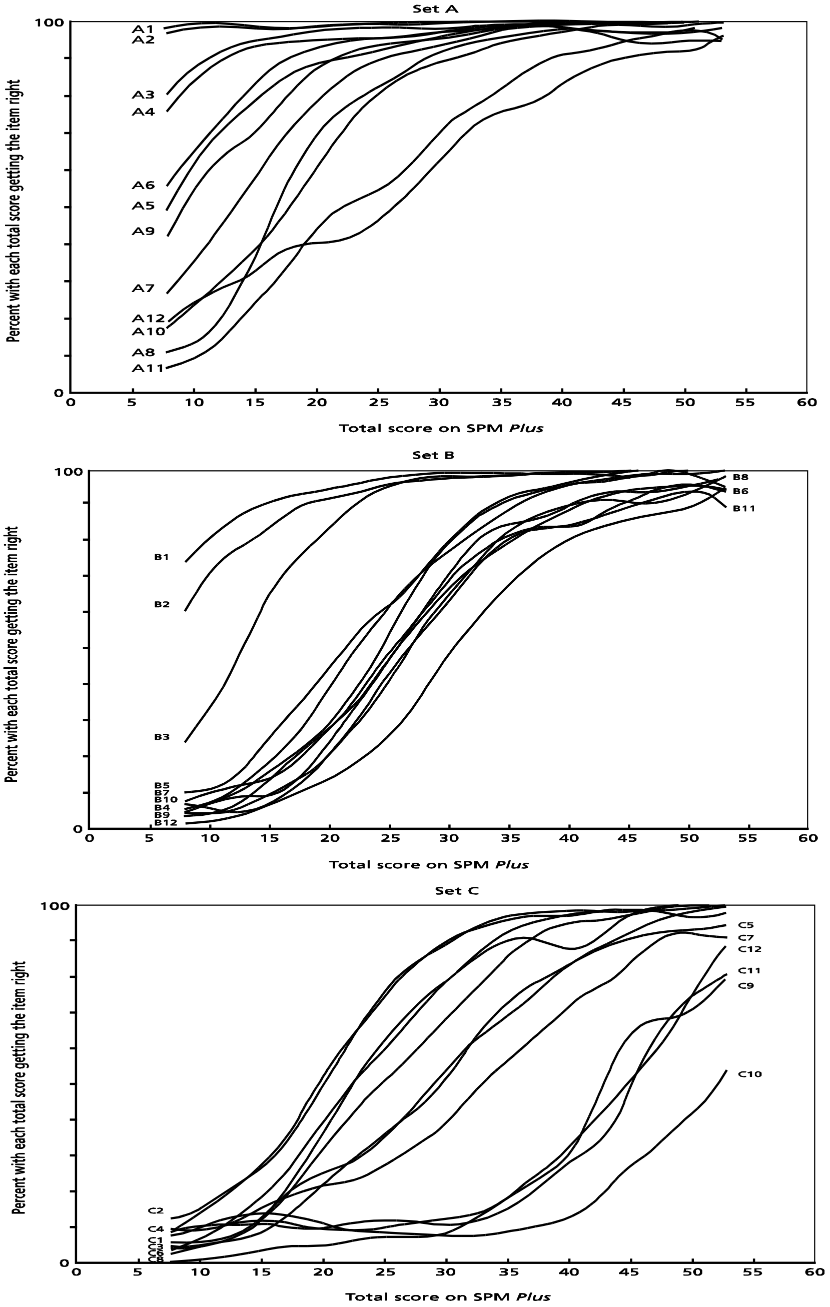
Graphing, Smoothing, and Transforming

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

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

Figure 5.4. *Standard Progressive Matrices* ***Plus*** – Romanian Standardisation

Empirical Item Characteristic Curves for Items Comprising Sets A to E (Smoothed)

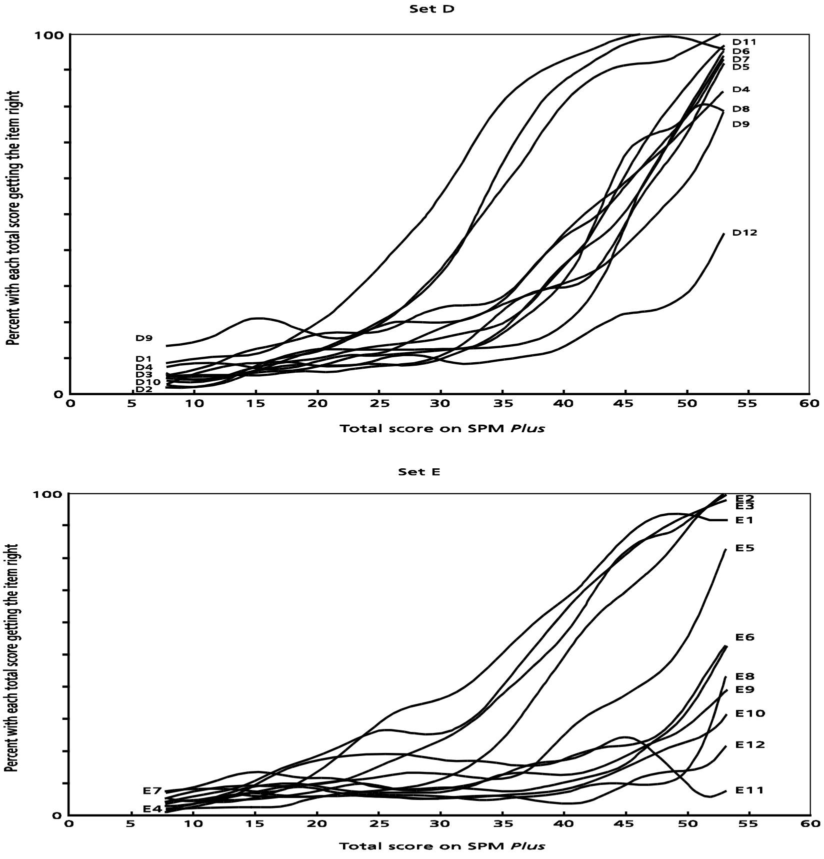


Set A

Set B

Set C

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



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

One point should perhaps be re-iterated here. Fischer's "reproduction" of the procedure originally employed when drawing the graphs by hand *smooths* the data. As we shall see, even the few mathematical-index oriented, computer based, IRT programmes that plot ICCs of the form published in Figure 2.4 in Hambleton, Swaminathan, and Rogers (1991) *transform* the data on the basis of one variant of mathematical IRT (e.g., concerning the shapes and slopes of the graphs) *before* plotting them. The mathematical indices outputted by these programmes are also "contaminated" in exactly the same way. They fail to reveal the "raw truth" about the items. Thus they do *not* enable one to follow the recommendations the APA task force on statistical inference (APA, 1999), which encourage researchers to examine plots of their raw data before deploying "sophisticated" statistical programmes.

Some Implications of the Fischer-Prieler "Empirical" ICCs

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

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

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

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

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

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

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

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

A 3-Parameter IRT Analysis

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

Two questions now arise:

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

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

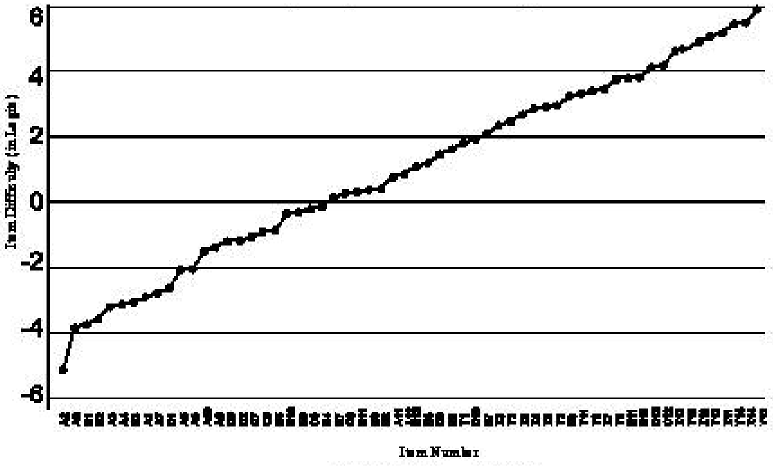


Figure 5.6. *Standard Progressive Matrices* ***Plus***
Romanian Standardisation
One-Parameter Model Item Difficulties Arranged in Order of Difficulty

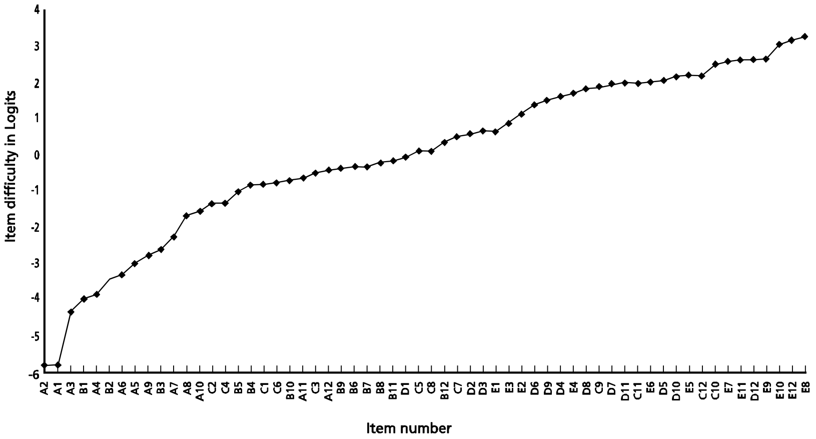
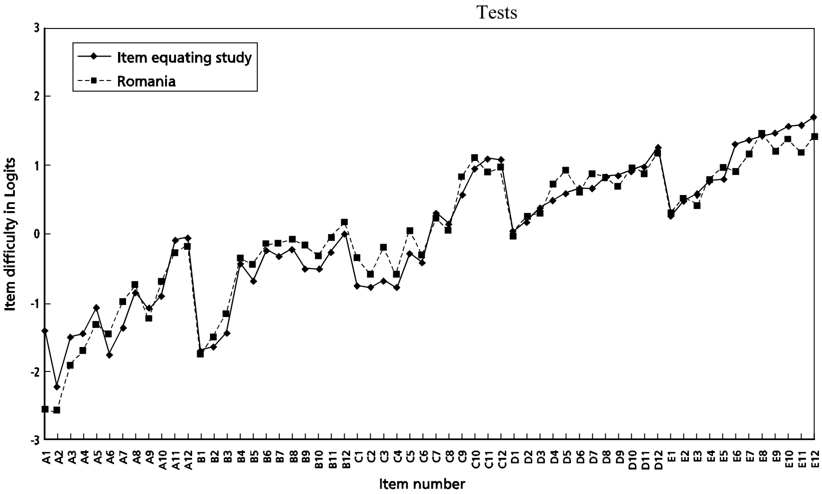


Figure 5.7. *Standard Progressive Matrices* ***Plus***
Comparison of Item Difficulties as Established in Item-Equating and Romanian Studies
One-Parameter Model Item Difficulties with Items Arranged in the Order in Which They Appear in the Tests



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

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

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

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

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

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

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

Notes:

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

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

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

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

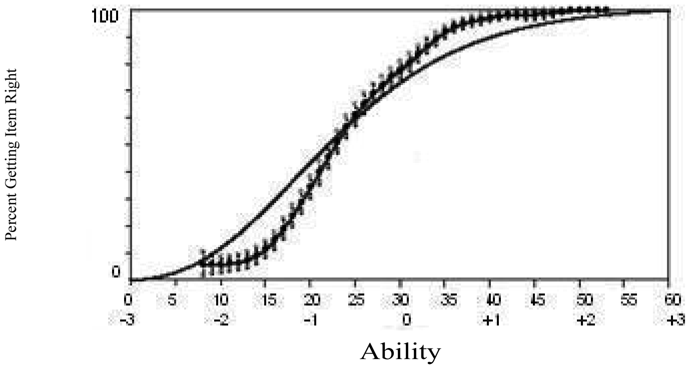
In Figure 5.10 the curve generated (with great difficulty) by fitting a 3-parameter model to the same data has been super-imposed onto the comparison of the empirical and 1-parameter ICCs for item C1 shown in Figure 5.8. Figure 5.11 presents similar comparative curves for Item D6. It will be seen that,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

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

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

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

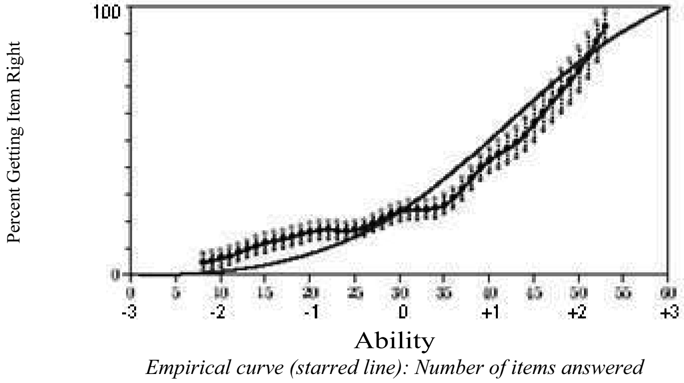
Figure 5.8. *Standard Progressive Matrices* ***Plus***
Romanian Standardisation
Comparison of Empirical and 1-Parameter ICCs for Item C1



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

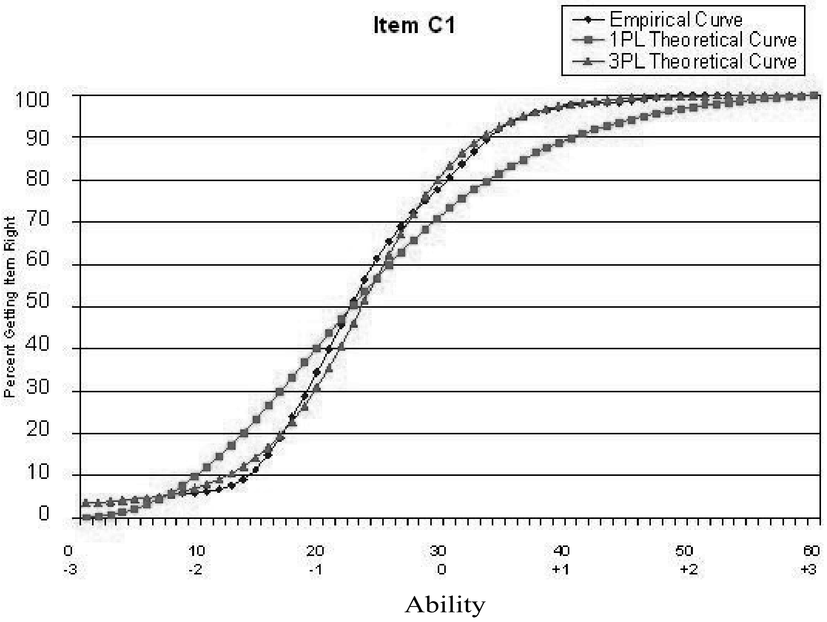
*IRT ability estimates are calculated from the difficulty and discriminative power of all items operational at each ability level.

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



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

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

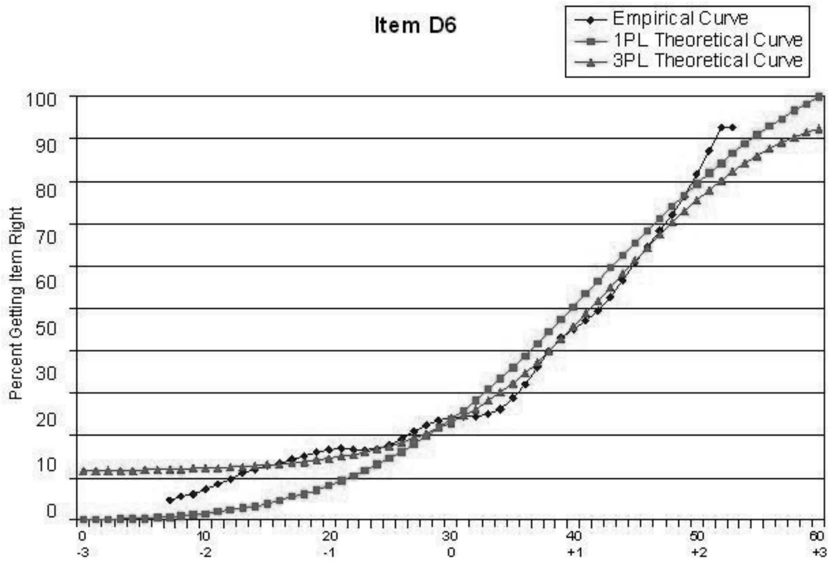


Empirical curve: Number of items answered correctly, i.e. total score on test.
1- and 3-Parameter IRT curves: Ability estimates in logits

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

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

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



Ability

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

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

Figures 5.12 and 5.13 show the plots of the ICCs of all items derived from the 1 and 3-parameter models. Nothing could give a better impression of the difference between the conclusions that follow from forcing the data into these alternative models. When the data are forced into a 1-parameter model, the ICCs appear to be of the same shape and evenly spaced. This is, presumably, a result of having employed item statistics derived from fitting a 1-parameter model to the data collected in the course of the item-equating and development study to select the items that were actually retained in the test. But the plot of the 3-parameter curves look very different indeed. The items are not equally discriminating; the order of difficulty varies with the ability of those taking the test, the items are not equally spaced, and they do not probe the domain of ability to be sampled by the test anything like as well as the plot of 1-parameter ICCs would have us believe. It is almost certain that, had we had sets of 3-parameter graphs generated from the data collected during the item-equating and development study we would have modified item A12 -- which is the item whose ICC crosses those for all the other items in the test. It is apparent that many of the least able respondents get it right before they have the ability to solve it and quite a number of the most able still get it wrong. Clearly, something about the item is distracting these able respondents.

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

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

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

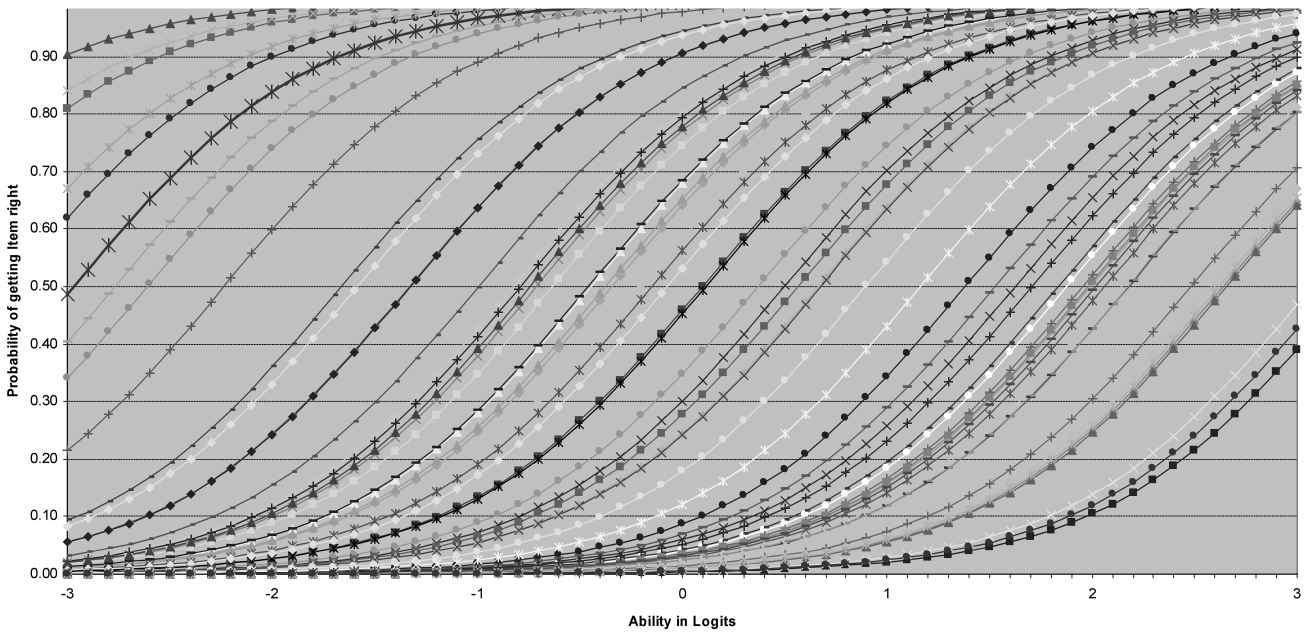
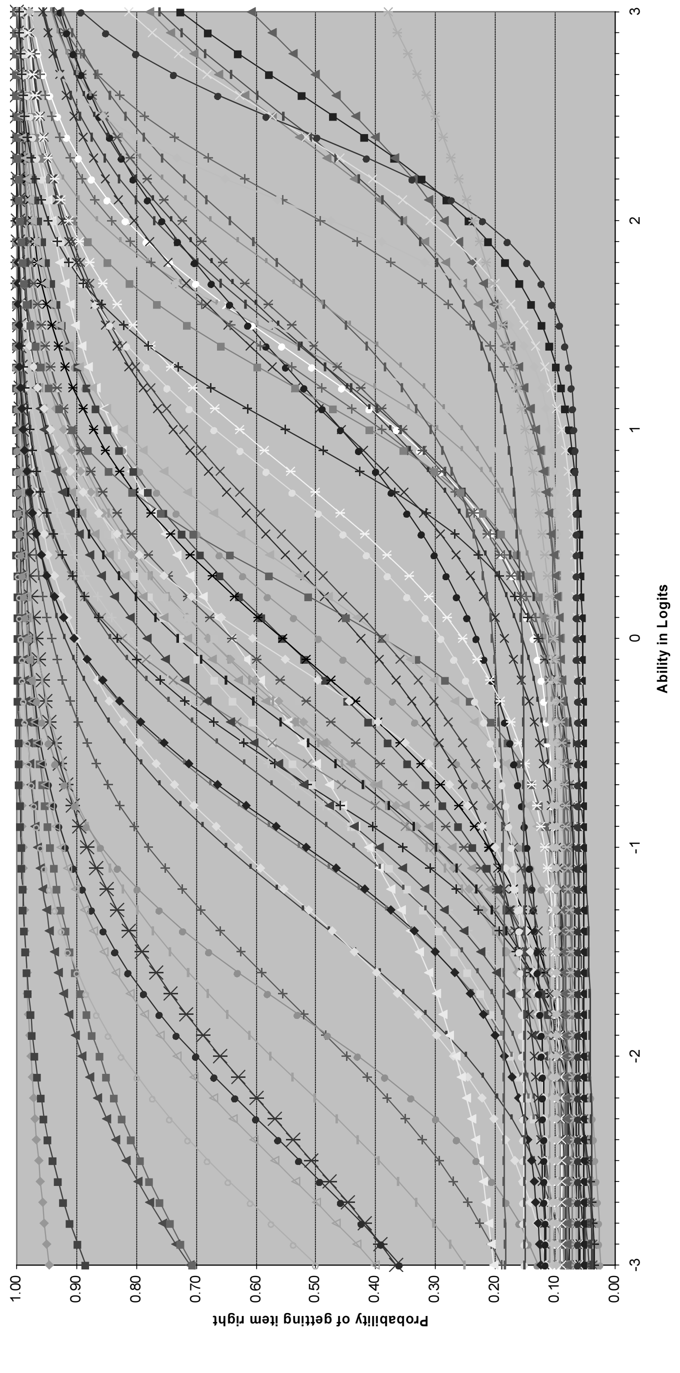
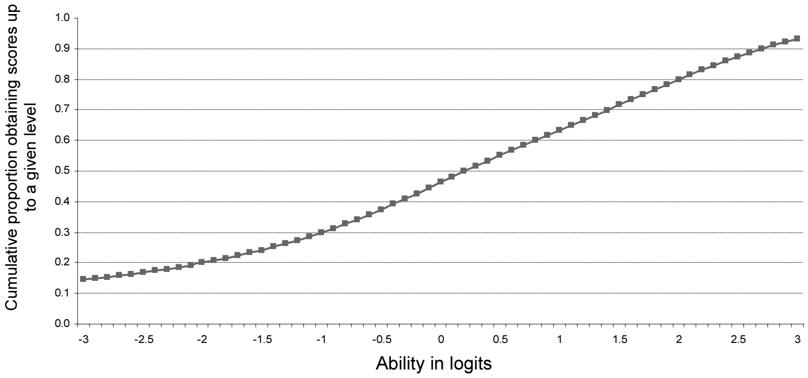


Figure 5.13. *Standard Progressive Matrices* ***Plus***
Romanian Standardisation
3-Parameter Model Item Characteristic Curves
(Each graph represents one Item)



Test Characteristic Curve and Test Information Function Curves

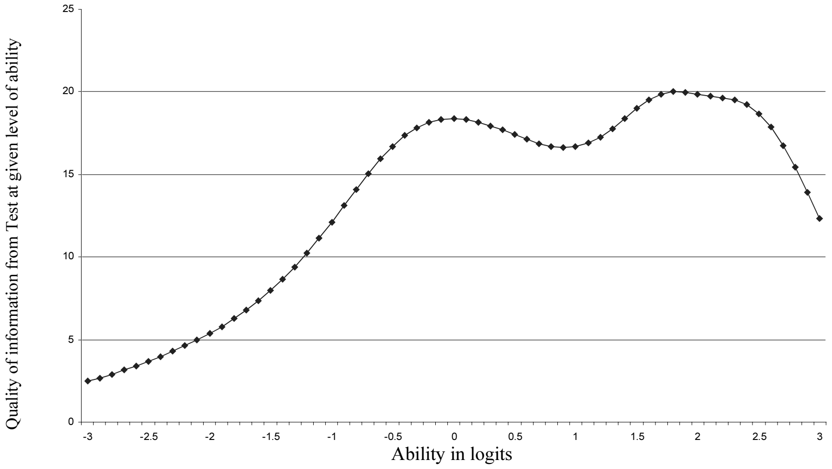
Figure 5.14. *Standard Progressive Matrices* ***Plus***
Romanian Standardisation
Test Characteristic Curve for 3-PL



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

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

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



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

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

Some Conclusions

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

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

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

Notes

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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