The Raven Progressive Matrices Tests: Their Theoretical Basis and Measurement Model

Article · January 2008

CITATIONS
45

READS
134,159

1 author:

John Raven

243 PUBLICATIONS **11,980** CITATIONS

SEE PROFILE

Chapter 1

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

John Raven*

Introduction

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

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

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

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

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

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

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

The Raven Progressive Matrices Tests and Their Philosophy

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

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

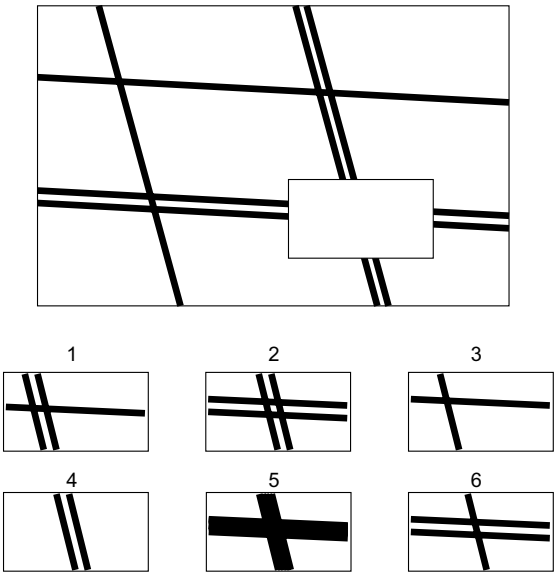
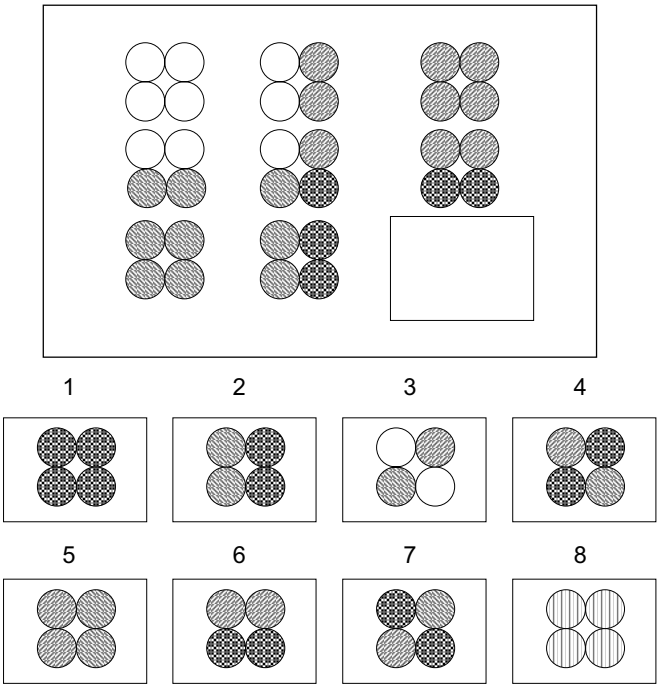
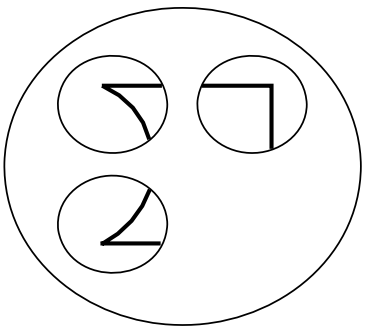


Figure 1.2 **A Moderately Difficult *Standard Progressive Matrices* Item**
(similar to one in the Test itself)



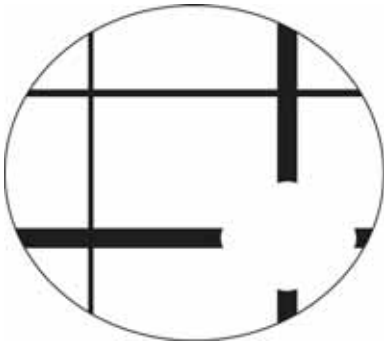
But, to illustrate what the tests are really trying to measure, it is useful to consider the "simpler" items shown in the next four Figures.

Figure 1.3



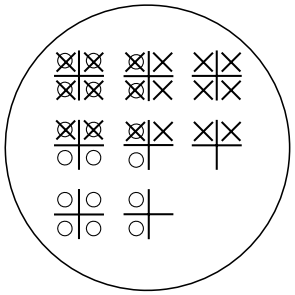
WHAT?

Figure 1.4



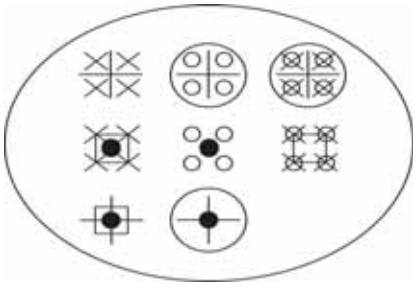
WHAT?

Figure 1.5



WHAT?

Figure 1.6

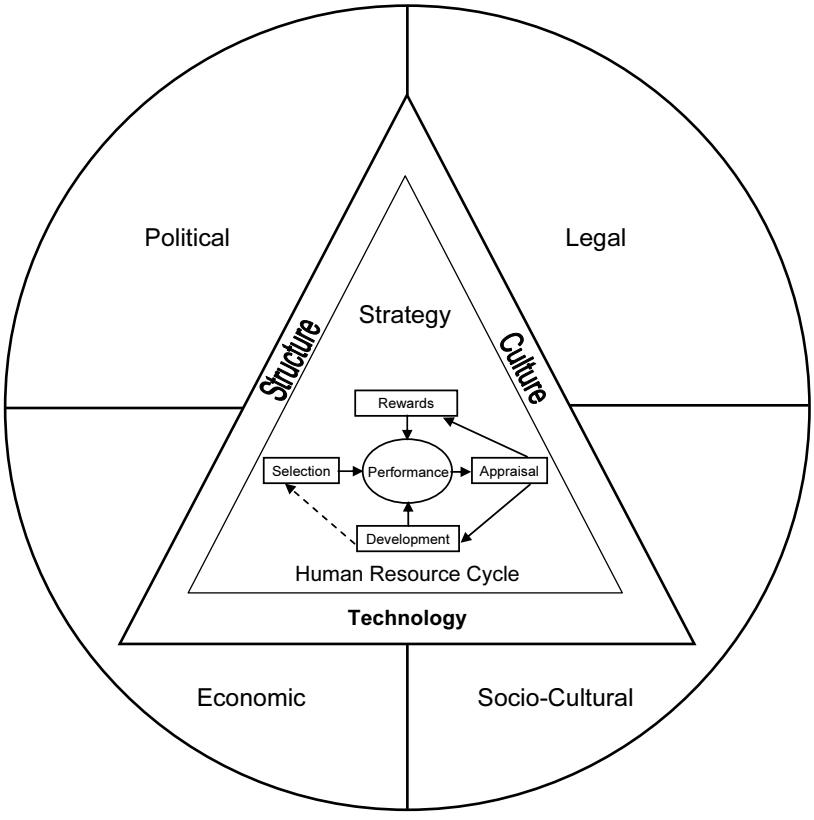


WHAT?

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

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

Figure 1.7



WHAT?

Reproduced, with permission, from Lees (1996)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

He made the point as follows:

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

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

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

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

The Measurement Model

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

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

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

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

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

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

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

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

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

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

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

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

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

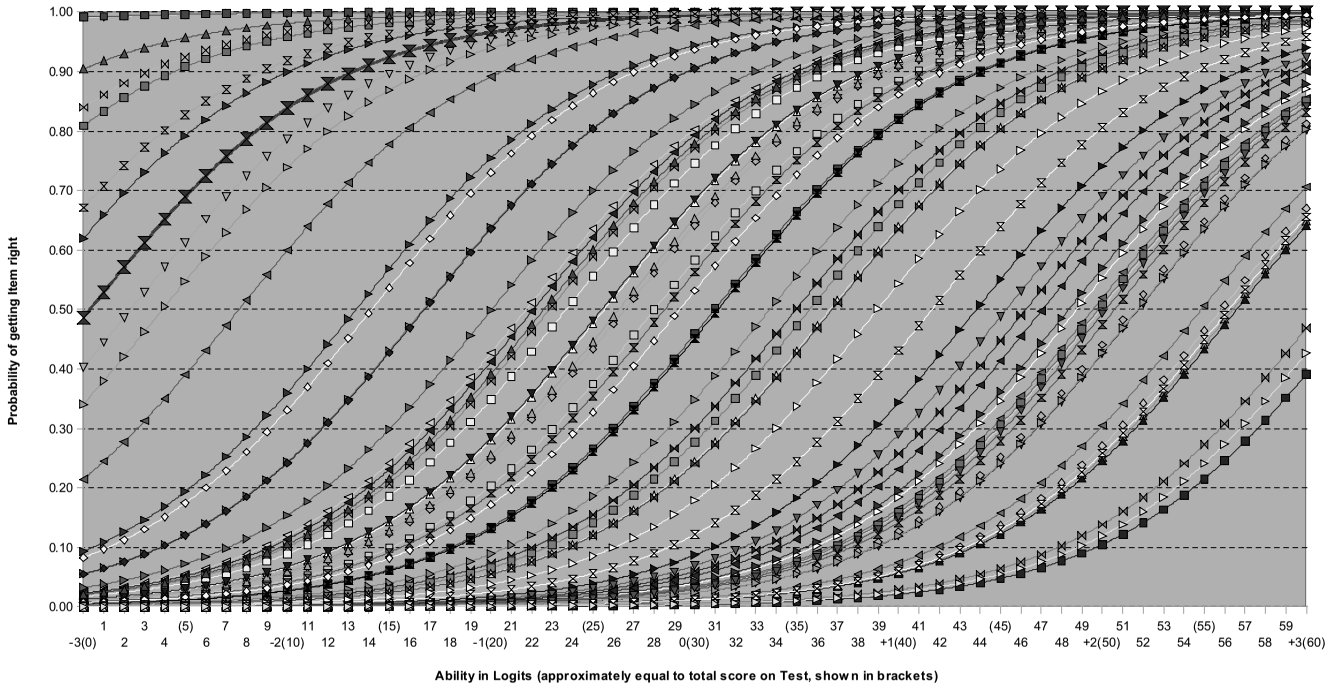
Thus, when the bar is set at very low levels – for example at the levels illustrated by the top first three curves (counting downwards) to intersect with the vertical axis – almost everyone, even of the lowest ability, is able to jump over it. But some of those with the lowest ability do knock it off.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

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

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

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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

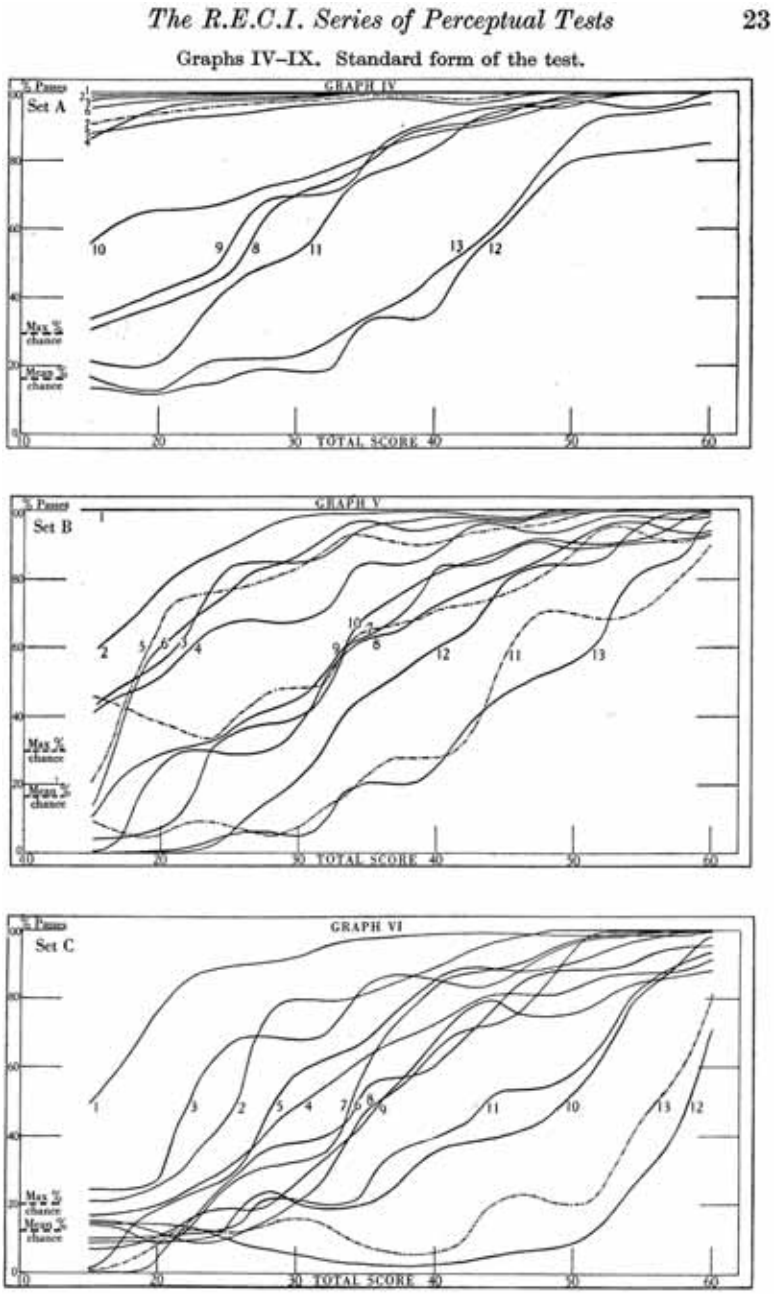
We hope that most readers will now be clear how radically the procedures used in the development of the RPM differ from those used in the construction of most other psychological tests. With the development of computers, the hand drawn graphing procedures used in the original development of the RPM have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

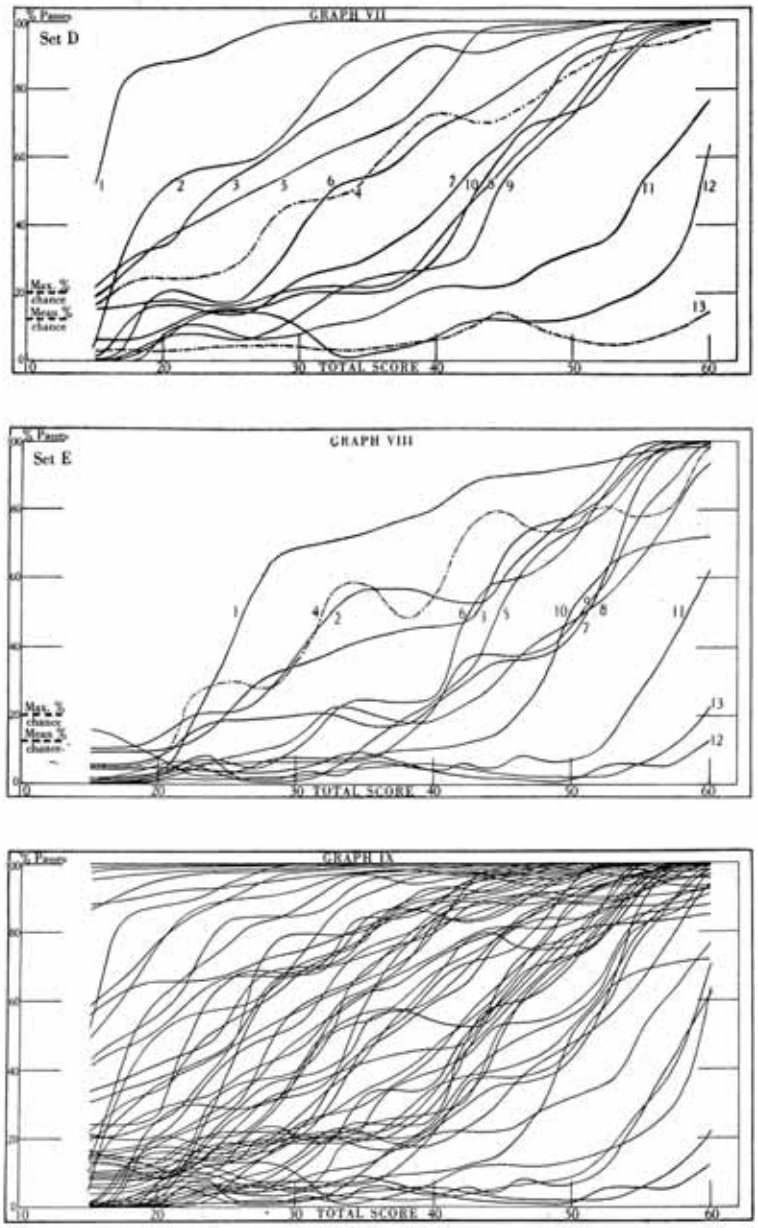
Unfortunately, because what has been said here was not so clearly articulated when the tests were first developed, few researchers understood the basis on which the tests had been constructed and this has led numerous researchers to draw misleading conclusions from their data … indeed to think in a confused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

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

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

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





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

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

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

This has two important implications:

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

The Robustness of the Measure

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

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

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

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

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

History of Test Development

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

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

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

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

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

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

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

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

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

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

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

Construct Validity: Another Preliminary Statement

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

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

Versions of the *Progressive Matrices* Tests

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

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

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

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

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

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

Some Findings

Heritability and the environment

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

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

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

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

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

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

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

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

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

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

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

Changes in Scores Over Time

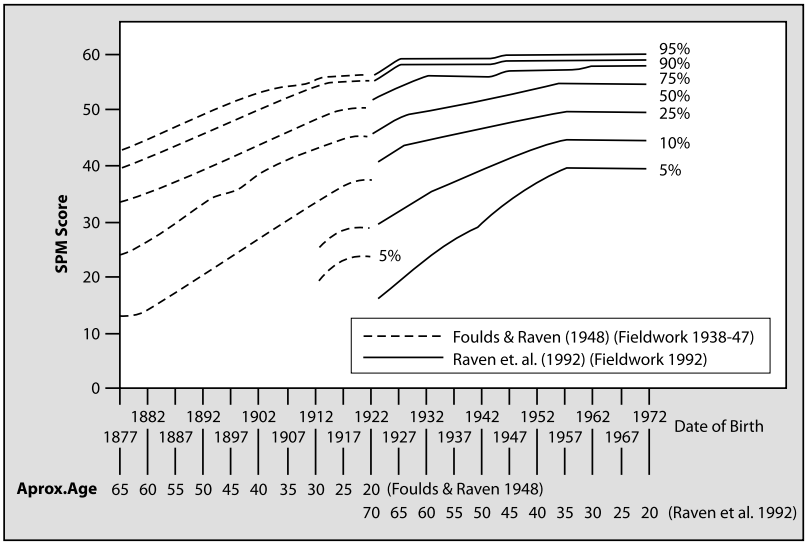
The most striking demonstration of the truth of Messick's statement so far as the RPM is concerned is to be found in research documenting huge inter-generational increases in scores[1.38].

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

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

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

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



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

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

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

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

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

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

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

A number of things follow from this.

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

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

The significance of this finding cannot be overestimated.

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

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

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

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

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

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

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

Stability in Norms Across Cultures

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

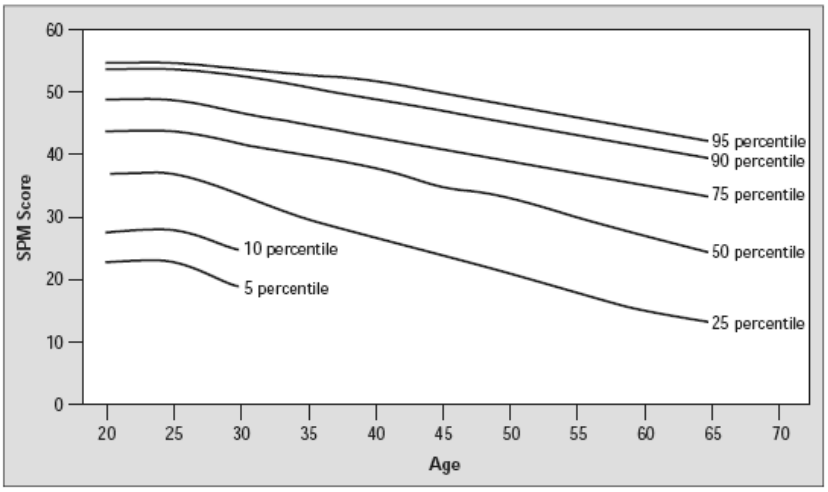
In the course of our introduction to this chapter we used graphical methods to show that the items of the RPM are not merely of steadily increasing difficulty but form a scale whereby the abilities required to solve any one item are systematically related to the abilities required to solve others (of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

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

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

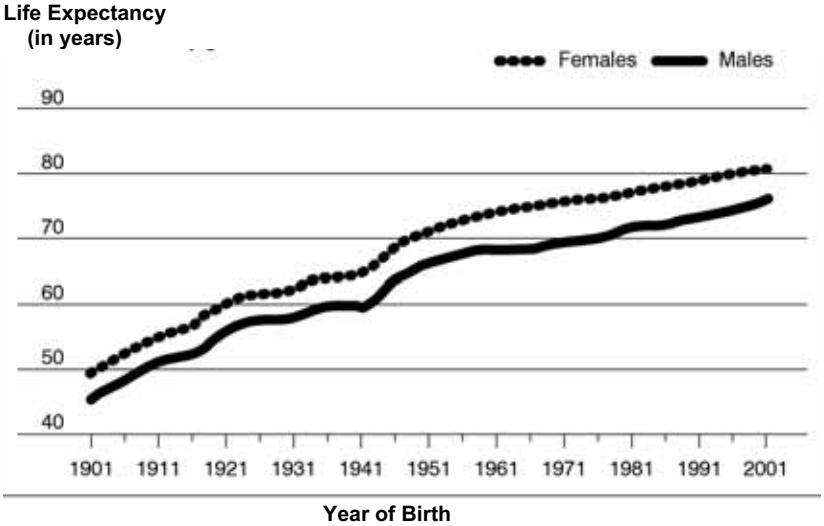
These data clearly support our earlier claim that the tests work in the same way – measure the same thing – in a wide range of cultural, socio-

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



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

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



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

Cross-cultural similarity in norms.

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

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

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

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

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

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

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

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

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

The Occupational Predictive Validity of the RPM

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Back to construct validity

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

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

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

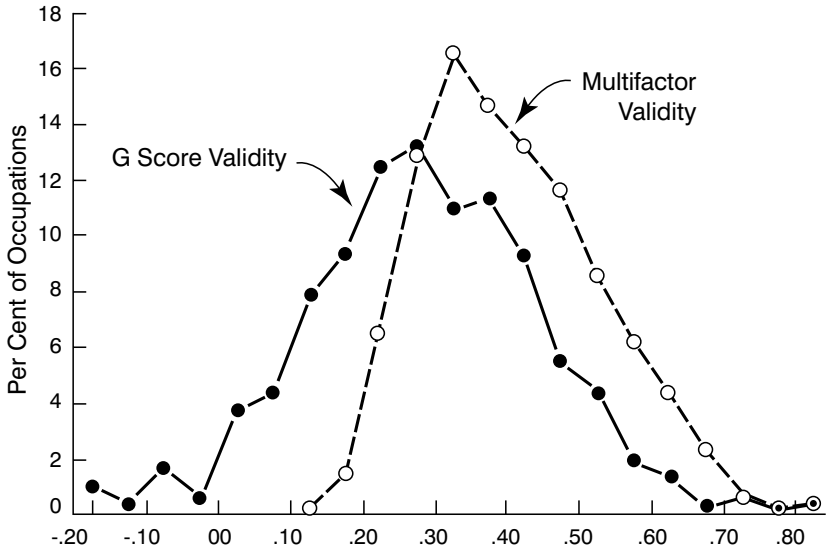
Note the problems which these observations pose for the validation and interpretation of "intelligence" tests: We have seen that children from similar backgrounds, including members of the same family, vary enormously both in their motives and values and their "intelligence". The variance in their motives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

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

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

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

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



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

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

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

More Recent Applications: Derived Scores

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

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

Notes

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

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

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

1.4 Kanter (1985)

1.5 Gardner (1991)

1.6 Spearman (1926)

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

1.8 Dewey (1910); Schon (1983)

1.9 Spearman (1926)

1.10 Messick (1995)

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

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

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

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

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

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

1.16 Vodegel-Matzen (1994)

1.17 Richardson (1991)

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

1.19 Spearman (1926, 1927a&b)

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

1.21 Deary & Stough (1996)

1.22 Cattell (1963)

1.23 Horn (1968)

1.24 Carroll (1993)

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

1.26 Cattell (1963)

1.27 Horn (1968)

1.28 Snow (1989)

1.29 Carroll (1993)

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

1.31 Horn (1994)

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

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

1.34 See, Maxwell (1961).

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

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

1.37 Messick (1989)

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

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

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

1.41 Raven (2000b)

1.42 Bouveir (1969)

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

1.44 Raven (1981)

1.45 Flynn (1987)

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

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

1.48 Garfinkel & Thorndike (1976)

1.49 Flynn (1999, 2000)

1.50 Flynn (1989)

1.51 Raven (1981)

1.52 Raven (2000a)

1.53 Jensen (1974)

1.54 Owen (1992)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.56 Eysenck (1953)

1.57 Schmidt & Hunter (1998)

1.58 Gottfredson (1997b)

1.59 Jensen (1998)

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

1.61 Hope (1984)

1.62 Kanter (1985)

1.63 Kohn & Schooler (1978)

1.64 Raven (1976, 1977)

1.65 Kohn (1977)

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

1.67 Kinsey, Pomeroy, & Martin (1948)

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

1.69 Sperry (1983

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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