Issues raised in developing and using a test of intelligence

Uses and Abuses of Intelligence: Studies advancing Spearman and Raven's Quest for Non-Arbitrary Metrics, John, Jean Raven (Eds.), 2008, Royal Fireworks Press, New York, ISBN 978-0-89624-356-7, 593 pp.

Introductory comments

In the 1930s John C. Raven developed the Raven Progressive Matrices (RPM) tests as a device for measuring Spearman's concept of *eductive ability*, something most of us today would call inductive reasoning or general reasoning. Raven's effort was successful. The tests, which include the Standard Matrices (now in a revised form, the SPM+), the Colored Matrices (for children) and the Advanced Progressive Matrices, are among the best known instruments for studying intelligence in use today. Arthur Jensen (1998), the modern-day standard bearer for Spearman's approach to intelligence, has pronounced them an excellent marker for general intelligence (*g*). This book, edited by John Raven (the son of John C. Raven, and a prolific publisher on psychology in his own right) and Jean Raven, contains 26 chapters, over half by John Raven as sole or co-author. According to the statement on the book jacket, the book succeeds in 'bringing together a series of studies stemming from Spearman's research on human abilities.' As an added feature, the book includes four chapters, two by Jim Flynn, that discuss broader issues.

Reviewing a collection of articles is always difficult, because of the variety of the articles and, often, the unevenness of the chapters. In this case the appropriate approach is made difficult because the book contains so many chapters by the same author, which makes it reasonable to expect more coherence between chapters than would be the case for the typical collection of efforts. But that is not the case. The flock of articles the Ravens have assembled are not all birds of the same feather.

I will first classify the papers, and explain why some of them will not be discussed at all. I then reverse the usual 'good news, bad news' presentation. I will present some criticisms of the book, as a book, that I think must be mentioned, but are not directly addressed to the intellectual issues raised. I then proceed to a critique of some major conceptual issues that are presented in several of the chapters in which Raven was involved.

Classification of the chapters

The chapters fall into the following classes:

I. Reports of national normative studies conducted in Romania, Slovenia, Lithuania, Turkey, Kuwait, South Africa, Pakistan, and nine Indian tribal groups (out of over one hundred different tribal groups in that diverse country, let alone the major ethnic groups.) There are also some chapters by Raven and his colleagues that discuss technical issues that were raised during the Romanian standardization. The information in these chapters belongs in technical manuals. It would be useful to have a single archive of all the national norming studies using the Progressive Matrices. (This could be another collection of chapters, but with a single theme, and would expand beyond the countries listed here.) The studies were well done, and will not be further considered.
II. Empirical studies using one or another of the matrices tests (and occasionally the Mill Hill vocabulary test). These range from an article by Irene Styles on the relation of Piagetian tasks to the RPM tests to a study of the relation between the test scores and driving records of truck drivers. The Styles article, which has not been previously published, is highly recommended. The others do not form a coherent theme, except that a common test was used, and properly belong in empirically oriented journals. Indeed, several of them were previously published and, in my opinion, were not sufficiently major to warrant re-publication here.
III. Chapters that raise philosophical or social issues going far beyond the use of the test. These include a case analysis of trials involving issues of mitigation of capital crimes due to diminished mental capacity and Flynn's précis of two of his books. There is a previously published discussion by John Raven, dealing with a wide range of social problems somewhat loosely related to testing. I can conceive of a carefully edited collection of papers on these topics, especially if they addressed common themes, but I see little use for three or four disconnected papers in a book that is primarily about tests and test interpretation.
IV. We then come to some of the interesting issues raised by John Raven and his colleagues, packed into three chapters; chapters 1 and 8 by John Raven, and Chapter 7, by Prieler and Raven. The same issues are raised in more than one chapter, and chapters make little reference each other. Therefore I will concentrate on the arguments themselves.

An unkindness of Raven's

But first I bring the bad news.

In the Elizabethan age the collective term for ravens was an "unkindness," which gives rise to my title for this section. This book was poorly edited. Some of the editing errors are just annoying, but others make the book much harder to read than it should be. Raven has done an unkindness to his readers.

Part of the fault lies with both the editors and the publishers. There is no index of authors or topics. Some of the figures are close to illegible. I had to use a magnifying glass at times. There is excessive repetition, both of arguments and exact duplication of figures. Some repetition is inevitable in a collection of papers, because different authors may say the same thing. Editors may hesitate to ask for revisions of previously published work that they, the editors, did not write. But that is not the case here. Most of the repetitions, including figures, are in papers that John Raven himself wrote. Surely he could have edited his own writing. This could have been done without cost to any of his arguments.

In addition to being repetitious across chapters, the argument is often poorly focused within chapters. The Prieler

and Raven chapter is a good example of the problem, but not the only one. This chapter makes reasonable points, but it makes too many of them, in an unfocused fashion. The book would have far more impact if Raven had distinguished between major and minor points, and focused on the major ones.

Raven is a wordy writer. Many of his arguments could have been made in fewer words. This is a serious problem in a book that stretches to just under 600 pages. I believe that had the book been edited severely, both in terms of writing and to eliminate repetitions the work would have been perhaps 450 pages long, and would receive much wider dissemination.

I found Raven's style of argumentation annoying. He often attacks the ideas of others, but gives no citation. Classic test theorists bear most of the brunt of charges that they do not understand something (usually related to IRT), but we are never told who the "classical test theorists" are or what they said that was so wrong. Similar vague charges are made on a few other topics. Argument by innuendo is more suitable to political than to scientific discourse.

These features of the book are annoying, but they are superficial. I close with a discussion of John Raven's view of an important issue that affects the interpretation of results using the RPM and many other tests as well. This is the point where I will attempt to play Chomsky's role, and use the book as a starting point for a broader discussion.

The role of scaling in studies of intelligence

The critique is written in the spirit of Naom Chomsky's (1959) famous review of Skinner's *Verbal Behavior.* Chomsky took Skinner's views as a starting point to raise broad issues about the study of language. Raven's writing, and especially his many criticisms of the work of others, generally revolve around the utility of using Item Response Theory (IRT) to construct psychological measurements, a practice he strongly advocates. I will discuss the pros and cons of relying so heavily on IRT in the study of intelligence. My critique is not meant as a sign of disrespect for John Raven. You know you have given a good colloquium when you are confronted with piercing questions. You know that you have given a bad colloquium when you receive a polite patter of applause, and then the audience goes home. Raven should take my review in that spirit. The critique is the good news.

Spearman, and following him J. C. Raven and now John Raven, wanted to study 'eductive ability.' John Raven documents the interesting historical fact that J. C. Raven reasoned that if eductive reasoning ability exist as a continuum, and if answering test questions requires eductive reasoning at some level of difficulty, *and nothing else*, then the probability that a person answers a given question correctly should rise as a function of that person's total score on a test containing similar items. The same principle underlies the better-known Guttman scaling technique. Subsequently the Danish psychometrician Georg Rasch developed a mathematical model that captures this intuition. Rasch's work is the basis of modern item response theory (IRT). John Raven documents how IRT is used to select items in the modern development of the RPM tests. He also attacks what he sees as misuses or misinterpretation of test scores. In order to understand the attacks it is necessary to consider what IRT does.

IRT is based on the assumptions that there is an underlying trait, here 'eductive ability,' that examinees can be ordered by the extent to which they possess this trait, and that test items can be ordered by the extent to which they demand the trait. Following common practice, I will refer examinee ability (θ) and item difficulty (β) parameters. The key insight is that ability and difficulty are measured on the same scale.

The probability that a person will answer an item correctly is assumed to be a logistic function of the difference between the person's ability and the item's difficulty. Loosely, if ability exceeds difficulty ($\theta > \beta$) the probability of getting the item correct is greater than .5, rising strictly monotonically to arbitrarily close to 1. Conversely, if difficulty exceeds ability the probability of getting an item correct falls toward zero. An important point is that the model assumes that equal differences between the θ and β parameters are assumed to have equivalent effects on the probability of getting an item correct, regardless of the level of these parameters.

This is called the *equal interval* assumption. To see how it works consider the following example. Suppose we have two items, an easy one with difficulty parameter β_e and a harder one with difficulty parameter β_h, where $\beta_e < \beta_h$. Suppose further that we have two examinees, a bright one with ability parameter θ_h and a not-so-bright one with θ_e. Item response theory requires that if $(\theta_e - \beta_e) = (\theta_h - \beta_h)$ the probability that the bright person gets the hard item correct must be equal to the probability that the not-so-bright person correctly answers the easy item.

In order to construct an IRT-compatible scale a test containing candidate items is given to a sample from a *reference population*. For example, the reference sample for the British standardization of the Standard Progressive Matrices was constructed from a sample of a British city chosen to be representative of the nation as a whole on a number of demographic variables. Through mildly involved statistical procedures θ and β values are found such that the observed response patterns make probability of passing a logistic function of the $\theta - \beta$ difference. This condition implies the equal interval assumption. During the process of test construction it may turn out that the responses to some items do not fit the logistic model. These items are dropped from consideration on the grounds that they 'scale improperly.' (Keep this in mind, it will be important shortly.) For the moment, though, consider what we can do with the test, once the β parameters are established.

Suppose that the new test, containing only properly scaled items, is given to a new sample, from a *test population*. We assume that the β parameters are fixed at the values established by the reference population. If we compute the θ values for the test population we have a way of comparing the test population and the reference population, on a scale that satisfies the equal interval principle.

Why not just compare these populations on a simple score, like total number correct? The answer to this question was

given over fifty years ago by S.S. Stevens (1957), and codified in a particularly lucid chapter by Suppes and Zinnes (1963). Their work, and related work by Duncan Luce and a number of others, introduced the concepts of *fundamental measurement* to psychology. Measurements of anything are expressed in one of several permissible scales. A frequently used example is temperature, which can be expressed on the Kelvin, Celsius, or Fahrenheit scales. These scales, like the IRT latent trait, satisfy the equal interval assumption. A statement about a set of measurements is *formally meaningful* if and only if its truth value is unchanged across all permissible scales. For instance, in the winter the mean temperature in Minnepolis is in the 20s, on the Fahrenheit scale, while Honolulu is in the 80s, on the same scale. It would not be formally meaningful to say that Honolulu is four times as warm as Minneapolis, because the truth value of that statement would be changed if we shifted to the Celsius or Kelvin scales. On the other hand, it would be formally meaningful to say that the range of temperatures from winter to summer in Minneapolis is equal to the range of temperatures in Moscow, Russia, or six times the range in Honolulu, because that is a statement about intervals. More generally, a scale satisfying the equal-interval assumption can be rewritten into another by a linear transformation. For that reason these scales are usually referred to as *linear scales*, although the equal-interval term does occur in the literature.

More generally, if a variable is measured on a linear scale then a statement about the measurements is formally meaningful if and only if the truth value of the statement is the same over all linear transformations of the scale. This is an important point, because the purpose of a scientific investigation is often to make a formally meaningful statement about some underlying variable, such as intelligence, whose value should not change if one permissible scale is substituted for another. For instance, we would not want a statement about the relation between a test score and, say, income later in life, to depend on whether or not the test was scored using the typical IQ scale (mean 100, standard deviation 50), T-scores, or standard scores.

In the case of linear scales statements about product-moment correlations and ratios of intervals (which include comparisons of the size of intervals, using F and t tests) are formally meaningful. However these statements are not formally meaningful over non-linear transformations. In particular, they are not formally meaningful over monotonic (ordinal) transformations of the scale. There's the rub. If a test is composed of items that fit the IRT model statements about correlations and intervals based on a conventional scoring method, such as number correct or percent correct, are not linear transformations of the underlying trait. Therefore statements comparing intervals or involving correlations could be true for measures of the underlying trait but not true for the conventional scoring or vice versa. At this point, and with these conclusions in mind, we turn to Raven's strongly expressed concerns about certain conclusions in the literature.

Raven argues, and a person committed to fundamental measurement theory at all costs would agree, that any comparison of the difference between two differences at varying absolute levels of ability should be made in terms of the underlying ability scale, not one of the typical conventional scoring methods. This is a relevant point whenever a claim is made that some treatment has a different effect at two different points on the intelligence scale.

Let us take an important case, which Raven cites repeatedly. Scores on standard progressive matrices tests have risen over the past fifty years. Using the "number correct" conventional scoring system the difference between the 10th percentile scores over this interval is greater than the difference between the 90th percentile scores. This has been interpreted to mean that the rise in intelligence observed in the last half century has been greater at the bottom than at the top. Is this justified?

Raven argues that it is not. He makes two arguments for his conclusion. One is that the tests involve are not sensitive at the upper end, therefore what we are seeing is a ceiling effect that artifactually depresses the change at the top. While it is true that this could happen, some of the studies showing differential gains have taken steps to guard against ceiling effects (Teasdale & Owen, 1989, 2008). In addition, If we compare the standardizations of the RPM tests over the years (illustrated here in several figures) we find that there is a steady decrease in the size of the increase in scores as the percentile increases. This is hard to account for solely by ceiling effects.

The chapter by Prieler and Raven makes a second, deeper argument. The cohort effect has been demonstrated by conventional scoring methods. The interpretation of the effect, though, refers to the underlying trait, ('eductive') reasoning ability. Because the conventional score is a monotonic non-linear transformation of the latent trait, one cannot claim that the gains at the top are smaller than the gains at the bottom, for that is a statement about the size of intervals. Such a statement is not formally meaningful for ordinal scales.

By the tenets of fundamental measurement this argument is correct. It could easily be the case that in the conventional scoring system the changes in the 10th percentile score over time could be greater than the changes in the 90th percentile score, although the changes on the underlying latent trait, 'eductive reasoning', were identical. As Prieler and Raven correctly note, the same argument could be applied to a training program that, according to an analysis of conventional scores, was more effective at the lower than the higher percentiles. The argument could be applied to a contrast between two groups, for example the British and Romanian data. If you demand that the IRT criteria be satisfied, then we could have inequality at different points on the conventional scoring scale and equality on the underlying scale, or vice versa. Of course, this argument applies to any test that is constructed to fit the IRT model, not just the RPM tests.

"Could be" is not the same as "is." The question of where the change is can easily be answered if one has access to the original (item level) data; you just determine where the θ values lie on the scale of the original test. In the case where the task is to evaluate changes after an intervention, Raven cites an even simpler procedure that depends on changes in responses to items. This technique requires that you have access to item level data on the same individuals, before and after the intervention.

But what about the case where you only have access to summary level statistics, e.g. the conventional scores for different percentiles of the 1942 standardization? Raven does not say what to do, but the problem is manageable. If the data fit the requirements of the simplest variation of the logistic model, the one-parameter logistic model (1PL), the number of correct items is a sufficient statistic for estimating θ. Unfortunately there is no way to calculate the standard error of the estimate, but for large samples this should not be a major problem. Raven could have made this comparison, but he did not. Thus he missed a chance to answer the question he correctly identified; has the change over time been greater at the bottom or the top of the intelligence scale?

Raven does not raise a problem that follows from the logic of his argument. He appeals to principles of fundamental measurement (albeit not by that name) to argue that we cannot interpret studies that depend upon comparisons of differences in conventional RPM scores. He is right, if you accept the principles of fundamental measurement. But, by exactly the same argument, one ought to regard all studies that depend upon correlations involving conventional RPM scores as uninterpretable. One can apply Raven's argument to conclude that the studies that lead Jensen to praise the RPM tests as a good measure of *g* were, in fact, uninterpretable. I do not think he wants us to reach this conclusion. Nor do I think we have to, but before explaining why I shift my focus to a second argument that Raven makes…once again in several places. As in the case of the linear scaling argument, his claims and concerns raise interesting issues that go well beyond the use of the RPM. I discuss this before presenting my analysis of the claims concerning linear scales, because my argument against Raven' conclusions about scaling is related to my argument concerning the other claims.

Raven's claim has two parts. The first is that the correct way to select items for a test (and illustrating by the construction of the Standard Progressive Matrices +) is to find items that fit the test format and produce item characteristic curves that conform to the logistic model. It turns out that substantial weeding of potential items is needed in order to get a test that has the desired mathematical properties. The second part of the claim is that because such a test retains its psychometric properties over several cross-cultural contexts, the trait underlying the test must be a near-universal characteristic of human reasoning.

Raven rests his argument for the cross-cultural claim on the fact that the various standardization studies for Romania, Latvia, etc. were successful. He also cites some published work and informal reports of very high correlations between item difficulty parameters in several different standardization samples. This is good evidence for Raven's contention that the test measures "the same thing" in different countries, within the industrial and post-industrial nations.

I would find the claim more impressive if the item selection had not been so pronounced. Step back and think about what has happened. A set of items is found that have certain psychometric properties when used to evaluate population A. These items have generally the same properties in population B and C. It has also been shown that the scale as a whole has important predictive value in population A; i.e. it does predict important performance, ranging from academic achievement to automobile driving records. That is impressive. It is also putative evidence that the scale might predict achievement in populations B and C, but that depends upon how similar the cognitive skills required in populations B and C are to those required in A. When A is one European country (usually Britain) and the other is a generally similar country (e.g. the United States) the demonstration of common scaling is enough to create a strong presumption of common predictive value. To the extent that countries B and C are culturally very different from each other (e.g. rural India) I would like to see both a demonstration of common scaling and a demonstration of common predictive value. Either an affirmation or failure of predictive value would be interesting, but depending on what happened one might draw different conclusions about the universality and utility of, to use Spearman's term, eductive reasoning.

I also have a concern about the emphasis on scaling as the *sine qua non* of item selection. For instance, Raven makes the point that items generated to conform to Carpenter, Just, and Shell's (1990) model of matrix test performance do not scale well. (He also says that a "noted expert" did not design good items, another annoying example of anonymous accusations. Either name names or keep quiet!) His argument seems to be that if items generated by a coherent psychological theory do not meet certain psychometric criteria then the theory must be wrong.

Rational people (including me!) may see the problem as more complicated. The psychometric tail should not wag the cognitive dog. To illustrate, Carpenter et al. showed that the performance on items that fit their model were systematically related to working memory capacity. It would be ideal if such items also fit a convenient psychometric scoring system, here IRT. If they do not, why not is a topic for further investigation, not a justification for ignoring the mathematically aberrant measures.

This brings me to a general question about the use of IRT and similar models as major criteria in test development. Make no mistake about it, I prefer and applaud the use of IRT and known scales, when it is possible to do so. But when it is not possible to do so, it is not clear to me that conformity to a mathematical abstraction is the defining property of a good psychological test. Raven's discussion of uses of the RPM tests brings the issue to the fore, but the issue goes well beyond the RPM tests.

Psychometric measures are generally developed gradually. Testing companies, for instance, regularly try out new items…as have Raven and their colleagues in development of the SPM+. The trial items are selected if they fit into the mathematical model (either factorial or IRT) that describes the response distributions obtained using the original test. Stripped of mathematical arguments, over time the test becomes a progressively better instrument for measuring whatever trait the original test measured. This has obvious advantages. But there is a devil lurking in the garden! What I have just described is a positive feedback mechanism, which any engineer knows is

a bad thing. Excessive adherence to this item selection procedure chokes off development of measures of psychological variability that go beyond the original depiction, when the test was first offered. To take the case at hand, John Raven's decision to reject items that do not scale well, regardless of their justification in psychological theory, produces better and better measures of Spearman and J.C. Raven's original conception of inductive reasoning but chokes off attempts to expand on that conception. This would be fine if the original conception was psychological truth. I think it was a pretty good conception, but I do not want to treat it as revealed truth.

Mathematically, IRT scaling does produce a linear scale. However it is an unusual example of such a scale. The zero point and scale of linear scales are usually produced by some reference to properties of the thing being measured. For instance, the Celsius scale of measuring temperature was generated by considering the temperatures at which an important substance, water, changes from a solid to a liquid, and then from a liquid to a gas. These points were chosen on the basis of a well established theory of what heat is and how it interacts with materials; in other words, the parameters of the scale were related to properties of the thing being measured.

IRT scales have the same mathematical properties as do the Celsius and Fahrenheit scales. However IRT parameters are not established by any theory of the thing being measured. In fact, the IRT scale for a test could itself be a linear combination of some more primitive traits, such as verbal and perceptual reasoning. The zero point and unit interval are established by the variability of whatever is being measured, something only defined by the items on the test, in the reference population. It is not at all clear to me that the rigorous demand for formal meaningfulness appropriate for measuring such things as temperature should be applied to statements about such amorphously defined psychological traits.

While it would be nice to use IRT scaling in psychometrics where possible, it is quite possible to make progress using conventional methods. The psychometrician Jum Nunnally (1978) reached this conclusion thirty years ago. While Nunnally applauded the use of fundamental measurement where possible, he did not feel that one should disregard studies where common-sense conventional scoring had been used. He presented a pragmatic and a scientific argument for his position. The pragmatic one is that although it is possible to construct non-linear, monotonic transformations that alter correlations drastically, the non-linearities must be extreme. To illustrate, the correlation between the integers 1…10 and their squares is .95; as is the correlation between these integers and their logarithms. In the case of the RPM and similar tests, if actual IRT scoring is not possible one might consider how extreme the distortion introduced by standard scaling has to be to make any difference to one's conclusion. And then ask yourself if such a distortion is likely to have taken place.

Nunnally's scientific argument was more telling. Many, many reliable results have been obtained with standard scoring methods. The observations are reliable and often have important theoretical and practical implications. The business of science is explaining these observations, not using measurement theory to explain them away.

Carried to its logical extreme, John Raven's arguments would have us either recompute a century of important results or, when this is not possible (as it usually will not be), throw these results away. Raven does not say this, he confines his use of the logic of measurement theory to attack some findings about differences between high and low scores. He does not take on the correlational studies. Nevertheless, if you believe that it is imperative to follow his advice concerning intervals you are on the first step down a slippery slope when you deal with correlations. We want to go there with caution.

References

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

Chomsky, N. (1959). Review of B.F. Skinner, *Verbal Behavior*. *Language*, *35*, 26–57.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT, US: Praeger Publishers/Greenwood Publishing Group.

Nunnally, J. (1978). *Psychometric methods* (2nd ed.). *New York: McGraw-Hill.*

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*, 153–181.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*, *Vol. I* . New York: Wiley.

Teasdale, T. W., & Owen, D. R. (1989). Continuing secular increases in intelligence and a stable prevalence of high intelligence levels. *Intelligence*, *13*(3), 255–262.

Teasdale, T. W., & Owen, D. R. (2008). Secular declines in cognitive test scores: A reversal of the Flynn effect. *Intelligence*, *36*(2), 121–126.

Earl Hunt
Psychology Department, Box 351525, University of Washington, Seattle, WA 98195-1525, United States
Email address: ehunt@U.washington.edu.

23 September 2008

doi:10.1016/j.intell.2008.09.005

The global bell curve: Race, IQ, and inequality worldwide, Richard Lynn, Washington Summit Publishers, Augusta, GA, USA, ISBN: 978-1-59368-028-2 (pbk) Pages: xviii, 298 pp. body text, 360 including references

This book is well organized and easily accessible to the generally educated reader. Like most of Richard Lynn's work, it reflects a relatively thorough and careful compilation of the relevant extant literature. The book begins where Herrnstein and Murray (1994) left off in *The Bell Curve*, with the observation that there is a socioeconomic hierarchy of race in the United States that can be attributed to intelligence test scores. It examines the degree to which this observation can be extended to other multiracial societies throughout the world that also show racial inequalities in earnings and socioeconomic status.