

Part XI: Intelligence: Defining Through Factor Analysis

18. Jensen's *g*: Outmoded Theories and Unconquered Frontiers*

PETER H. SCHÖNEMANN

ON RELEVANCE

From the total literature spanning more than a century, a few 'bad apples' have been handpicked most aptly to serve Gould's purpose. Yet what relevance to current issues in mental testing are the inadequacies and errors of early anatomical studies. . . . Who now wishes to resurrect . . . Terman's pronouncements in 1916 about eugenic measures to reduce the incidence of mental retardation; the primitive 1916 Army mental tests; or the US Congress's 1924 Immigration Restriction Act, which cited the 1917 Army test data? (Jensen, 1982a, p. 124)

. . . haven't all sciences always exercised free license for theoretical speculation about the causes of their observable phenomena in their domains? Of course they have. (Jensen, 1982a, p. 131)

In (1969) Jensen availed himself of this license to speculate whether 'current welfare policies, unaided by eugenic foresight, could lead to the genetic enslavement of a substantial segment of our population?' (p. 95). He expressed concern that 'the possible consequences of our failure seriously to study these questions may well be viewed by future generations as our society's greatest injustice to Negro Americans' (*ibid.*).

His concern for the future welfare of this particular minority group had apparently been aroused by the findings of a number of surveys aimed at evaluating the effectiveness of the compensatory education programs of the early 1960s. It is true that most of them 'produced absolutely no significant improvement [effect] in the scholastic achievements of disadvantaged students' (*ibid.*, p. 3). However, it is also

*I would like to thank Professor Jensen for his generous response to my request for reprints of his work. Drs Cicirelli, Mayeskee and Wang for help with the literature search, Mr Dorsey for bringing Alan Chase to my attention, and my teachers, Drs Amthauer, Cattell, Cronbach, Humphreys, Kaiser, Lord and Tucker for introducing me to the theory and practice of mental test theory.

true that the designs and analyses of these hastily thrown together evaluation studies have been debated ever since. It now appears they tell us more about the intellectual limitations of the researchers than of the subjects. Less controversial as a fact, though no less ambiguous in its implications, is Jensen's observation that, 'on the average, Negroes test about 1 standard deviation (15 IQ points) below the average of the white population in IQ.' (*ibid.*, p. 81). Such bare facts can mean many things. For example: 'Unfortunately, not all children in our society are reared under conditions that even approach the optimal in terms of psychological development. One socially significant result of this is the lowering of the educational potential of such children' (Jensen, 1966, p. 238).

By 1969 Jensen felt this explanation required revision. After reviewing a number of empirical studies, the concept of the IQ, 'the nature of intelligence', and 'the genetic basis of individual differences in intelligence in humans', he concluded: 'There will be greater rewards for all concerned if we further explore different types of abilities and modes of learning, and seek to discover how these various abilities can serve the aim of education. This seems more promising than acting as though only one pattern of abilities, emphasizing *g*, can succeed educationally, and therefore trying to inculcate this one ability pattern in all children' (Jensen, 1969, p. 117).

These recommendations have stirred up much controversy, because it is hard to avoid the impression that they amount to a call for resegregation of the educational system which would perpetuate social injustices in the US. To appreciate how this impression could have arisen, one has to study some of the fine print: 'Level I ability is tapped mostly by tests such as digit memory, serial rote learning, selective trial-and-error learning with reinforcement (feedback) for correct responses, and in slightly less "pure" form by free recall of visually or verbally presented materials, and paired-associate learning' (*ibid.*, p. 111). The critical point here is that Level I ability is not 'intelligence' *per se*, but only a prerequisite for it. Much of this ability pattern can be found in rats and simulated with computers. 'Level II abilities, on the other hand, [are] ... best measured by intelligence tests with a low cultural loading and a high loading on *g*—for example, Raven's Progressive Matrices' (*ibid.*). This '*g*' ... can be regarded as the nuclear operational definition of intelligence, and when the term intelligence is used it should refer to *g*.' (*ibid.*, p. 9). No matter how Jensen goes about it, he invariably finds whites do better than blacks on Level II tasks. This leads him to conclude that 'the disadvantaged child's strongest point [is] the ability for associative learning' (*ibid.*, p. 115), i.e., Level I ability.

The study of the 'genetic basis of ... intelligence in humans', including work by 'the most distinguished exponent ... of these methods ... Sir Cyril Burt' (*ibid.*, p. 33), has further convinced him that either 75 or 76 per cent of the 'variance in IQ's is attributable ... to genetic variation', depending on how he estimates it. Few would quibble that there is 'quite good agreement between the two estimates' (*ibid.*, p. 51).

The rest follows by elementary logic. It would be futile to try to uplift 'the disadvantaged' from their innate Level I to Level II intelligence, which is intelligence in the technical sense of *g*. Federally funded compensatory education programs cannot 'boost IQ and scholastic achievements' because 'intelligence', *g*, is largely innate and 'the disadvantaged' are deficient in *g*-genes. While such efforts may produce a transient 'hot-house effect' (p. 103), in the long run they are 'to the disadvantage of many children whose mode of learning is predominantly associative',

i.e., who are non-intelligent in the sense of *g*. 'Accordingly, the ideal of equality of educational opportunity should not be interpreted as uniformity of facilities, instructional techniques, and educational aims for all children. Diversity rather than uniformity of approaches and aims would seem to be the key to making education rewarding for children of different patterns of ability.' (*ibid.*, p. 117).

Such a resegregation into Level I schools and Level II schools has the further advantage of minimizing, at least as a first step, the danger of 'possible dysgenic trends' which Jensen worries about (*ibid.*, p. 91). This brings us to the second step: suppose after some time Level I people are no longer needed to sweep the streets and man the gas stations, because robots can do this more cheaply. What do we then do with the Level I people? Jensen points the way: 'Have we thought sufficiently of the rights of children—of their rights ... not to have a retarded parent...? Can we reasonably and humanly oppose such rights of millions of children as yet not born?' (*ibid.*, p. 93).

Whatever else one may hold against Jensen, it cannot be said that his topic lacks social relevance. His writings show that he is aware of this. His retort to Gould suggests that he may be less aware of the historical relevance of this particular topic. This is surprising, because on other matters, e.g., the Stroop test (Jensen and Rohwer, 1966) and reaction time (Jensen, 1982b), Jensen has demonstrated an unusual degree of sensitivity to historical connections.

ON DEFINITIONS

Intelligence, like electricity is easier to measure than to define. (Jensen, 1969, p. 5)

Measurable intelligence is simply what the intelligence tests test, until further scientific observation allows us to extend the definition. (Boring, 1923)

What we measure with [intelligence] tests is not what the tests measure—... (Wechsler, 1975)

In truth, 'intelligence' has become a mere vocal sound, a word with so many meanings that finally it has none. (Spearman, 1927)

If we say, for instance, 'a submarine is a ship which can go under water' we define, not a submarine, but the term 'submarine'. (Reichenbach, 1947, 1975, p. 20).

The mental testers have more difficulty with little questions than with big questions. Jensen can tell us 'whether our collective intelligence is adequate to meet the growing needs of our increasingly complex industrial society' (Jensen, 1969, p. 88). He can also tell us with uncanny precision to what extent 'intelligence' is inherited: 75 or 76 per cent, depending on how he measures it (*ibid.*, p. 51). But he cannot tell us what he means by 'intelligence'.

When Jensen addresses the man on the street, he takes 'an operational stance. First of all, this means that probably the most important fact about intelligence is that we can measure it' (*ibid.*, p. 5). How does he know this? We learn that '... Spearman hypothesized the existence of a single factor common to all tests involving complex

mental processes [which] Spearman called "general intelligence" or simply *g*. It can be regarded as the nuclear operational definition of intelligence, and when the term intelligence is used it should refer to *g*' (*ibid.*, my emphasis).

A hypothesis is simply an unsubstantiated conjecture. Suppose that Spearman's hypothesis were wrong. What possible relevance could it then have for Jensen's operational definition? Suppose, on the other hand, that Spearman's conjecture were correct. Then it can be shown that it still is inadequate as an operational definition of 'intelligence' because Spearman's model defines not just one, but infinitely many *g*s for the same data. Confronted with a simple numerical demonstration (Schönemann, 1983), Jensen did not challenge the mathematics ('... we may take it for granted that it is [correct]', Jensen, 1983, p. 313). He simply dismissed it as 'wholly gratuitous and sophistic, at best' (*ibid.*).

Jensen is not alone with his vague allusions to *g* when it serves his purpose, only to discard it again when the going gets rough: 'We cannot enter into all this here, but only indicate our own position by saying that Spearman's generalized proof of the two-factor theory constitutes one of the great discoveries of psychology' (Wechsler, 1939, p. 6). One problem with Wechsler's interpretation of Spearman's work is that, in an empirical science, great discoveries do not require 'generalized proof', but empirical evidence. A few decades later Wechsler just as casually unloaded Spearman's 'proof' as a prop for his test again: '... the profusion of factors discovered seems to contradict the intent or purpose of the factorial technique.... Actually, there seem to be more factors than available tests, certainly good tests of intelligence' (Wechsler, 1958, p. 127).

What, then, is *g* all about which so intrigued Jensen and Wechsler? It is about an attempt to obtain an operational definition of 'intelligence' which, moreover, was to be quantitative, so that the degree of 'intelligence' could be expressed numerically. This was the claim implied by the programmatic title of Spearman's (1904) paper: "'General intelligence", objectively determined and measured'. To avoid confusion with the vague verbalisms of his predecessors, he later substituted 'general ability' ('*g*') for 'intelligence'. The quotes around 'General intelligence' make it clear that Spearman knew he defined a term, 'not a thing' (Jensen, 1983, p. 314). Many testers are confused about the purpose of a definition. It is not to give a whole theory about what is meant by a word, but just the intended meaning of it. The 'thing' comes in at a later stage: 'Definitions are arbitrary, and we may include whatever predicates we wish in a defined term; but after a definition is given there always remains the question whether there is a corresponding thing' (Reichenbach, 1947, 1975, p. 90). It is pointless to define a 'digoo' as 'an odd number divisible by 4' because no such thing exists. In an empirical science this means one has to produce some empirical evidence that the thing, as defined, exists in the real world before the definition can become useful. This is the hard part, and the more conditions one invests in a definition, the harder it will be to find a corresponding thing. Failure to acknowledge this is the major shortcoming of the many facile verbal 'definitions of intelligence' the testers have offered.

Most people want to use the term 'intelligence' to make statements of the type: 'this person is more intelligent than that person.' This implies that they think 'intelligence' can be expressed numerically, by just one number, so that the order relation among the numbers can be used to make inferences about the degree of intelligence of the subjects. This is not a necessary requirement for an operational definition. However, if it can be imposed, the practical usefulness of the definition will

be greatly enhanced, because then we can 'measure' the defined concept. The price to be paid for this added convenience is the added challenge to demonstrate that a 'thing' exists which satisfies the requirements of order, in particular, transitivity: whenever *A* is more intelligent than *B*, and *B* is more intelligent than *C*, then *A* must be more intelligent than *C*, however else one may have defined 'intelligence'. If we are unable to produce such a thing, then this definition of 'intelligence' is empty (has no empirical content), and we may discard it as useless.

Spearman also knew that measurement is equivalent to a numerical definition, an operational definition which involves a (structure preserving) numerical map. It is complete nonsense to claim to be able to 'measure intelligence' without being able to define it: 'Intelligence, like electricity is easier to measure than to define.' Such wayward examples from physics are often used by the mental testers to deflect attention from their own record. Instead of simply telling us what they mean by the term 'intelligence', they talk about thermometers, electricity and quantum theory. One of the few impressive achievements of the mental testers is to have succeeded in talking the general public into believing that it is possible to 'measure intelligence' without being able to define it.

Boring's definition of 'intelligence' comes closest to an operational definition of 'intelligence' as a measurable quantity, because an 'intelligence test' can be viewed as a map which associates a number (the IQ) with each person who has taken the test. The problem is that there are many different 'intelligence tests' and that they are imperfectly correlated. This means that *A* may be more intelligent than *B* on test 1, *B* more intelligent than *C* on test 2, while *C* is more intelligent than *A* on test 3, so that transitivity is violated. Hence, if one wanted to adopt Boring's definition, it would either have to be tied to a particular (standard) test—which is unlikely to be agreed on for commercial reasons—or one would have to qualify all statements about 'intelligence' with a reference to the particular test used. Jensen has employed this language occasionally, as when he described the results of some preschool intervention programs: 'The average IQ gains on three different tests were 5.32 (Peabody Picture Vocabulary), 2.62 (Stanford-Binet), and 9.27 (Wechsler Intelligence Scale for Children)' (Jensen, 1969, p. 105).

Such a strategy would permit the unambiguous evaluation of research claims about 'intelligence' at least in principle. However, it would also mandate that any research claim about 'intelligence' be validated for each IQ test separately. The psychometric notion of 'intelligence' rests on the hope that this may not be necessary because all three tests 'really measure the same thing', except for measurement error. It, therefore, rests on a very strong hypothesis which may well be false: the above three tests may measure different things, or nothing of interest at all. Moreover, if we are not careful how we formulate this hypothesis, it may no longer nail down 'the same thing' unambiguously.

It was precisely this problem Spearman proposed to solve with his Two Factor Theory. He observed that most 'intelligence tests' then in use correlated positively, but not perfectly, with each other ('positive manifold'). This he interpreted as an indication that the tests measured the same thing. Spearman argued that the reason for the lack of perfect correlations among the 'intelligence tests' was that they all measured not just one but two things, 'true intelligence', *g*, and 'error' (hence: 'Two Factor Theory'). The error attenuated the correlations among the observed 'intelligence tests'. He further argued that the error affecting any one observed 'intelligence test' should be uncorrelated with *g*, and also with the errors affecting all other

'intelligence tests'. Hence, if it were statistically possible to remove *g* from the observed scores, then we would be left with uncorrelated error. This prediction, he felt, provided an empirical test of his Two Factor Theory. It later became the basis for a methodology now known as 'factor analysis'.

Spearman spent considerable effort in developing and refining practical tests for his theory. On applying them to numerous 'intelligence' data which had already appeared in the literature, he found his prediction confirmed: after 'partialling out' just *one* latent variable, *g*, he found in each case that the residuals were indeed uncorrelated. This result seemed to confirm that there was a measurable thing deserving to be called 'intelligence' and thus validate his claim of having 'objectively determined and measured "intelligence"' or, as we now would say, of having operationally defined it. His definition was based on an empirical law: every 'intelligence score' consists of *g* and an error component which is uncorrelated with *g* and all other errors, and nothing else.

Had Spearman been successful, it would have been a momentous achievement. Among other things it would have spared Jensen the embarrassment of having to admit under pressure that he still does not know what he means by 'intelligence': 'the fact that the concept of intelligence is not clearly defined is not troublesome' (Jensen, 1983, p. 314). Maybe not to the mental testers—but what about the 'disadvantaged' who have already been sterilized on the basis of the flimsy scholarship of the intelligence experts?

ON MANIFOLDS

This doctrine was based upon what we have all along been finding of such paramount importance, namely, the correlations between abilities. (Spearman, 1927, p. 72)

... all positive correlations among all the diverse tests, known as a *positive manifold* ... is a central empirical fact for research on human mental abilities. (Jensen, 1983, p. 314)

One of the most striking and solidly established phenomena in all of psychology is the fact of ubiquitous positive correlations among all tests of mental ability.... This is simply a fact of nature. (Jensen, 1980, p. 249)

It should be stressed that Spearman's Two Factor Theory predicted more than just a positive manifold. It also predicted that the statistical removal of just *one* underlying variable, *g*, would reduce the observed correlation matrix to a diagonal correlation matrix of uncorrelated residuals. More importantly, the converse was to hold: if removal of just one variable reduced the correlation matrix to diagonal form, then we were supposedly entitled to the claim of having defined exactly one common factor, *g*, which qualifies as an operational definition of 'intelligence'. In this section we will take a closer look at this reasoning. We shall ask three questions:

- 1 How startling a 'fact of nature' really is the positive manifold, once we take into account how 'intelligence tests' are usually constructed?
- 2 How stringent is the inference from a positive manifold which satisfies Spearman's model to an underlying trait, such as *g*?

- 3 Suppose the positive manifold were indeed generated according to Spearman's model. Given the correlation matrix and observed test scores, what can we then learn about *g*? In particular, are we then entitled to make statements of the type 'this minority group has more *g*, on the average, than that minority group'?

THE GREAT SOCIETY

Two tribes live in the Great Society, the Alphas and the Betas. Both worship Greatness. If one says to an Alpha person, 'Gee, you look Great today', he will consider it a compliment. All types of Greatness are considered positive. Anything large, big, tall, long or numerous is viewed as a sign of Greatness and desirable: large families, high income, big houses, long names, tall stature, large shoe sizes, are all considered Great. Anything small, short or puny is frowned on. Both tribes are known for their skills in Pseudometrics, the science of measuring undefined variables (e.g., Bodenlos, *Ueber die Abzahlbarkeit des Nichts*, Pseudometriks, 1884; Listless, *Latent Structures of Empty Relations*, Pseudometriks, 1964).

While the Alphas and Betas agree on the desirability of Greatness, they do not agree on who is Greater. Since there are so many kinds of Greatness, it is not always clear how they should be compared. Three possible 'indicators' of Greatness are:

- Test A: number of great Aunts
- Test B: number of letters in the month of Birth
- Test S: Shoe size

After giving these tests to four randomly selected Alphas and four randomly selected Betas, the following Greatness scores were observed (above average: 1, below average: -1):

	A	B	S		A	B	S
Alpha 1	-1	-1	-1	Alpha means	0	0	0
Alpha 2	-1	1	1	Beta means	0	0	0
Alpha 3	1	1	-1	Total means	0	0	0
Alpha 4	1	-1	1	SDs	1	1	1
Beta 1	1	1	1				
Beta 2	1	-1	-1				
Beta 3	-1	-1	1				
Beta 4	-1	1	-1				

Alpha 4 has fifteen great aunts, which is far above average, and she wears shoes size 12, which is also above average. Hence she received standard score +1 on both tests. Since she was born in June, which has only four letters, she scored -1 on Greatness test B. Beta 1 has twelve great aunts (above average), wears shoes size 12½ and was born in December (eight letters). Hence he received a positive score on all three Greatness tests. At the right of the scores, the means and standard deviations (SDs) are given. Since the subgroup means for the Alphas and the Betas are zero for each Greatness variable, these data provide no evidence that the Alphas are Greater than the Betas on any of the three Greatness indicators.

Since all scores are standard scores (have mean zero and variance 1), one obtains

the correlation coefficient between any two variables, say test A and test B, simply by multiplying the score on test A with that on test B for each person and then dividing the sum of these products by the number of people ($N = 8$). On performing these simple computations for all possible pairs of tests, A with B, A with S and B with S, one obtains three correlation coefficients, r_{AB} , r_{AS} , and r_{BS} . They all turn out to be zero in this case. This means the three Greatness tests are perfectly uncorrelated, so that their correlation matrix is

$$\begin{array}{c|ccc} & A & B & S \\ \hline A & 1 & 0 & 0 \\ B & 0 & 1 & 0 \\ S & 0 & 0 & 1 \end{array} = I$$

This situation is reminiscent of Wissler's (1901) finding that the various 'cognitive tests' developed by the Wundt school (reaction time, visual acuity, speed of crossing out e's, etc.) were virtually uncorrelated with each other and also with school grades. Binet found later that much better results could be obtained with more complex, hodge-podge tests representing a broad sample of diverse abilities.

The pseudometricians also constructed three hodge-podge tests each of which sampled some of the three uncorrelated tests, an AS-Inventory (ASI), a BS-Scale (BSS) and a BA-Test (BAT):

$$ASI = .707A + .707S, \quad BSS = .707B + .707S, \quad BAT = .707B + .707A$$

The scores on these hodge-podge tests are:

	BAT	ASI	BSS		BAT	ASI	BSS
A1	-1.414	-1.414	-1.414	A-mns	0	0	0
A2	.0	.0	1.414	B-mns	0	0	0
A3	1.414	.0	.0	T-mns	0	0	0
A4	.0	1.414	.0	SDs	1	1	1
B1	1.414	1.414	1.414				
B2	.0	.0	-1.414				
B3	-1.414	.0	.0				
B4	.0	-1.414	.0				

The total means (T-mns) and the subgroup means (A-mns and B-mns) are again zero on all three tests. Hence, in terms of the 'manifest' test scores, there is again no difference in Greatness between the Alphas and the Betas on the hodge-podge tests. However, an Alpha pseudometrician noticed that the correlations among the ASI, BSS, and BAT, while not perfect, were all positive ('Great Manifold'):

$$\begin{array}{c|ccc} & BAT & ASI & BSS \\ \hline BAT & 1.0 & .5 & .5 \\ ASI & .5 & 1.0 & .5 \\ BSS & .5 & .5 & 1.0 \end{array} = R = Y'Y/N$$

In a paper entitled "General Greatness", objectively determined and measured' (Pseudometriks, 1904), he proposed a new Greatness theory which postulates that the

observed scores y_{ij} contain 'measurement error'. Hence the observed scores cannot be trusted as measures of 'true Greatness', G . Rather, each observed test score y_{ij} is a weighted average of a standardized true Greatness score, x_i , and a standardized, test-specific error score, z_{ij} :

$$y_{ij} = x_i a_j + z_{ij} u_j, \quad i = 1, 8; \quad j = 1, 3,$$

where the latent variables x and z_j have variance 1, and a_j, u_j , are the weights ('factor loadings'). The variable x is true Greatness in standard score form. If we define the vector $a' = (a_1, a_2, a_3)$, the diagonal matrix $U = \text{diagonal}(u_1, u_2, u_3)$, and the two score matrices $x = (x_i)$, $Z = (z_{ij})$, then the matrix of observed test scores $Y = (y_{ij})$ can be written

$$Y = xa' + ZU, \quad \text{var}(x) = 1, \quad \text{Var}(Z) = I_3, \quad \text{Cov}(x, Z) = \emptyset.$$

This Greatness model will be called the 'Great Model', for short.

What distinguishes the Great Model from other, earlier models and definitions of Greatness is that it can be tested: if this model holds, then the observed correlation matrix R among the manifest Greatness tests, ASI, BSS, and BAT must be of the form

$$R = aa' + U^2,$$

where U^2 is a diagonal matrix. Hence, on subtracting this diagonal matrix from R one obtains a 'reduced correlation matrix'

$$R - U^2 = aa'$$

which has exactly rank 1. Such a diagonal matrix U^2 is given by

$$U^2 = \begin{pmatrix} .5 & & \\ & .5 & \\ & & .5 \end{pmatrix}$$

which reduces R to the rank 1 matrix

$$\begin{pmatrix} .5 & .5 & .5 \\ .5 & .5 & .5 \\ .5 & .5 & .5 \end{pmatrix} = \begin{pmatrix} .707 \\ .707 \\ .707 \end{pmatrix} \begin{pmatrix} .707 & .707 & .707 \end{pmatrix}$$

$$R - U^2 = a \quad a'$$

The square-roots of the diagonal elements of U^2 give the weights for the standardized error variables, and the components a_j of the vector a the weights with which x , true Greatness, enters into each hodge-podge test. In the example, these weights are all equal to .707. The fact that we were able to compute these two weight matrices, U^2 and a , from the observed correlation matrix R without leaving any residuals means that Spearman's model holds exactly for these data. All that remains to be done is to find each subject's Greatness score, x_i . We will then be able to settle once and for all who is Greater, the Alphas or the Betas.

TRUE GREATNESS

This last step is a bit more problematic than the solution for the weights a_j, u_j . For a long time most pseudometricians believed 'Greatness scores cannot be computed, but

only estimated.' They devised numerous ingenious formulae for 'estimating' the true Greatness scores in some optimal way, e.g., 'in a least squares sense'. The regression weights for estimating true Greatness in a least squares sense are given by the vector

$b = R^{-1}a,$

according to these pseudometricians. Since the inverse of *R* is

$R^{-1} = \begin{pmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix} / 2 \text{ and } a = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} / \sqrt{2}$

one finds

$b' = (1 \ 1 \ 1) / 2\sqrt{2} = (.354 \ .354 \ .354)$

so that

$\hat{x}' = b'Y' = (-1.5 \ .5 \ .5 \ .5 \ | \ 1.5 \ -.5 \ -.5 \ -.5)$

are the 'least squares estimates' (LSEs) of the Greatness scores *x_i*. In terms of these estimates there is again no difference between the two tribes since both subgroup means are zero.

But these 'least squares estimates of Greatness' are not really true Greatness scores. For one thing, the Great Model says true Greatness scores have variance 1, whereas the Greatness estimates *x̂_i* have only variance *a'R⁻¹a* = ¾. For another, if we were to compute the 'least squares estimates' of the error scores *z_i* by the same logic, they would be correlated with each other and with *x̂*. According to the Great Model, they should be uncorrelated.

As it turned out, we do not have to settle for such imperfect Greatness estimates. A leading Alpha pseudometrician eventually succeeded in producing a set of true Greatness scores, *x_i*, together with an associated set of test-specific (error) scores, *z_{ij}*, which reproduce the observed test scores *y_{ij}* in exact agreement with the Great Model. These exact fitting true Greatness scores are given by:

	<i>x</i>	<i>z</i> ₁	<i>z</i> ₂	<i>z</i> ₃	(<i>x̂</i>)		<i>x</i>	<i>z</i> ₁	<i>z</i> ₂	<i>z</i> ₃	(<i>x̂</i>)
A1	-1	-1	-1	-1	(-1.5)	A-mns	.5	-.5	-.5	-.5	0
A2	1	-1	-1	1	(.5)	B-mns	-.5	.5	.5	.5	0
A3	1	1	-1	-1	(.5)	T-mns	0	0	0	0	0
A4	1	-1	1	-1	(.5)	SDs	1	1	1	1	.75
B1	1	1	1	1	(1.5)						
B2	-1	1	1	-1	(-.5)						
B3	-1	-1	1	1	(-.5)						
B4	-1	1	-1	1	(-.5)						

To illustrate, A1's BAT-score is *y*₁₁ = .707(-1) + .707(-1) = -1.414, A3's BBS-score is *y*₂₃ = .707(1) + .707(-1) = 0. The first column gives the true Greatness scores, *x_i*, and the last column gives their 'least squares estimates', *x̂_i*, for comparison.

Since the average of the true Greatness scores is .5 for the Alphas, and -.5 for the Betas, the Alphas are one full standard deviation Greater than the Betas, which is what the Alphas had been saying all along. In terms of GQs (Great Quotients, GQ_{*i*} = 100 + 16*x_i*), this is equivalent to an average GQ of 108 for the Alphas against an average GQ of 92 for the Betas.

One can well imagine how the Betas must have felt at the time of this discovery. It seemed so unfair to them since the observed test scores showed no mean difference whatsoever. The Beta pseudometricians spent countless hours of computer time searching for a more equitable Greatness solution until finally their efforts paid off. Not only were they able to produce another set of true Greatness scores which reproduced the observed scores *Y* exactly according to the Great Model, but their solution also confirmed their long-held conviction that they were truly Greater than the Alphas, once measurement error is taken into account. The Beta solution of the true Greatness problem is as follows:

	<i>x</i>	<i>z</i> ₁	<i>z</i> ₂	<i>z</i> ₃	(<i>x̂</i>)		<i>x</i>	<i>z</i> ₁	<i>z</i> ₂	<i>z</i> ₃	(<i>x̂</i>)
A1	-2	0	0	0	(-1.5)	A-mns	-.5	.5	.5	.5	0
A2	0	0	0	2	(.5)	B-mns	.5	-.5	-.5	-.5	0
A3	0	2	0	0	(.5)	T-mns	0	0	0	0	0
A4	0	0	2	0	(.5)	SDs	1	1	1	1	.75
B1	2	0	0	0	(1.5)						
B2	0	0	0	-2	(-.5)						
B3	0	-2	0	0	(-.5)						
B4	0	0	-2	0	(-.5)						

These scores also reproduce *Y* as *Y* = *xa'* + *ZU* in perfect harmony with the Great Model. Since the Betas have an average true Greatness score of .5 while the Alphas have an average Greatness score of only -.5, this solution puts the Betas one full standard deviation (16 GQ points) ahead of the Alphas.

On comparing these true Greatness scores (first column) with their 'least squares estimates' (last column), one finds that the estimates do not do justice to the Betas, because they consistently underestimate their true Greatness scores, while overestimating the true Greatness of the Alphas. This explains why there were no differences in the observed means, even though the Betas are Greater than the Alphas, once measurement error is taken into account. In the previous true Greatness table which favored the Alphas, the situation was reversed.

The Alphas and the Betas were not able to resolve all their differences, because each side tenaciously clung to their own true Greatness score solution. But they did agree that the 'least squares estimates' were no better than the observed scores in concealing the true Greatness difference between them. They found little solace when someone proved that these 'estimates', at least, whatever they may be estimating, are unique (Staggering, *Unique LSEs of Ducks in Flight*, Pseudometriks, 1967).

There only remained the problem which of the two tables of True Greatness scores was 'really true'. The correlation between the two sets of True Greatness scores in ⅔ = .5 in this case. This is not atypical for 'intelligence' tests, as long as the number of factors is small relative to the number of observed variables (see Schönemann and Wang, 1972, for minimum correlations based on various ability data, and Schönemann, 1981a, for an illustration of the indeterminacy problem in terms of Wechsler data).

ON UNCONQUERED FRONTIERS

As to the existence of 'g' as a common factor, there seems to be no possibility of doubt. Psychometrics, without it, loses its basic prop. (Wechsler, 1939, p. 8)

In order to keep the numerical illustration as simple as possible, it was stated in terms of 'exact sample data' and sample statistics. All the foregoing could have been stated for the population case, where random variables replace scores, expected values replace summation signs, and population parameters replace sample statistics. Thus the indeterminacy is not limited to 'factor scores' but applies with equal force to the random variables of the factor model, such as *g*.

A transposition into any variety of component analysis (Schönemann and Steiger, 1976), e.g., principal components, does not provide a way out, as Jensen seems to think, who wavers between factors and components. All linear composites of the observed variables (including the so-called 'factor score estimates') are entirely dependent on the particular variables in the battery and hence change from battery to battery. Similarly, Jensen's 'working definition of intelligence ... [as] *g*, or the first principal component of an indefinitely large and varied battery of mental tests' (Jensen, 1979, p. 17) does not suffice for an operational definition of 'intelligence' either. First, because *g* is not defined as a principal component, but as a factor. Second, because it is not clear what a principal component of an 'indefinitely large' correlation matrix is, since the largest eigenvalue tends to infinity. Third, because it is no longer an operational definition because we can administer only tests of finite length.

We now return to the three questions which were posed at the beginning of this section:

- 1 We have seen that it is not at all difficult to generate positive manifolds from complete nonsense data. All that is required is that we form several linear combinations from any given set of variables, which may well be entirely unrelated, so that the weights are all non-negative.

On reading any edition of *Buros' Mental Measurements Yearbook*, or by studying the contents of these tests directly, one finds that most IQ-tests are composed of very similar subtests like sentence completion, analogies, number series, same-opposites, arithmetic reasoning problems, memory tasks, in various admixtures. The total test scores which deliver the IQs are simply sums, i.e., non-negative linear combinations, over such subtest scores. It therefore proves very little if IQs of different tests, say the Wechsler and the Stanford-Binet, correlate highly with each other. Nor is it astonishing that the subtests of such hodge-podge tests produce a positive manifold. It would be more surprising if a subtest calling for logical inferences about numerical relations were entirely uncorrelated with a subtest which calls for such inferences about verbal or pictorial items, just as it would be surprising if left and right shoe size were entirely uncorrelated. In short, positive manifolds, which have been hailed as a 'fact of nature' and made the cornerstone of the claim that *g* must exist, can frequently be explained as simple artefacts of test-construction procedures.

- 2 There remains the fact that one can often account for positive manifolds in terms of just one latent variable. Thomson (1916) has shown that this result can also be produced by forming arbitrary non-negative linear combinations of *very many* uncorrelated variables. This means concretely that each test may require a very large number of mini-skills, rather than just one mental superskill. As the number of variables composing each observed variable approaches infinity, the observed correlations will fit Spearman's model

exactly (for more details and numerical examples see Schönemann, 1981b, p. 331). Thus, even when Spearman's model fits exactly, we need not infer it was generated by just one common factor. It could just as well have been generated by infinitely many mini-skills. On the basis of the data, we cannot decide how the observed variables were generated.

Thorndike advanced such a multifactorial theory of intelligence already in (1903). Intuitively, it is at least as plausible as Spearman's theory of a single factor *g*. The endless fragmentation of factors over the years, which even Wechsler lamented, has more recently turned into an equally endless parade of new *gs* (verbal *g*, numerical *g*, crystallized *g*, fluid *g*, true *g*, etc.). All this makes a multifactorial theory of 'intelligence' more plausible (see also Carroll and Horn, 1981; Tuddenham, 1962; and the 'componential' theorists) and negates the notion that it could be fair and realistic to summarize 'intelligence' with just one number, 'the IQ', or that it makes sense to speculate how much of this superskill is inherited.

Parenthetically, it should also be noted that, in assuming a perfect fit of Spearman's model, we ignored that no one really knows how to test these models with any stringency, in spite of the existence of numerous maximum likelihood algorithms which routinely deliver statistical tests of fit. The problem is that these tests tell us only the probability of falsely rejecting the model. Virtually nothing is known about the much more serious problem of falsely accepting it. (For more details on the consistent neglect of the power problem in maximum likelihood factor analysis see Guttman, 1977; Schönemann, 1981a, 1981b.)

- 3 We finally found that even when Spearman's model fits and we choose to ignore the Thomson/Thorndike interpretation, we would still be left with many different *gs*, because the latent variables of the factor model are indeterminate. We can assign many different scores to the people and obtain the same fit.

This so-called 'factor indeterminacy problem' was first pointed out by E. B. Wilson in 1928. Jensen (1983) believes 'this limitation of factor scores has been recognized by most other factor analysts' (p. 313). Be this as it may (see Steiger and Schönemann, 1976, for an uncensored history of this issue), it still remains to be explained why Jensen has never mentioned the indeterminacy issue in his many mini-tutorials on factor analysis before, given the significance it has for his belated attempts to resurrect Spearman's *g*. One reason may be that Jensen still misses the point. The problem is not, as he thinks, 'that factor scores derived from a common factor analysis are indeterminate in the sense that they are imperfectly correlated with the hypothetical factors' (Jensen, 1983, p. 313; I confess I do not understand what this means). Rather, the problem is that Spearman's *g*, which is the foundation of Jensen's operational definition of 'intelligence', is indeterminate. There are not just one but infinitely many 'intelligences' which all explain the same data equally well. The correlations among these many different 'intelligences' may be negligible.

Wechsler was perfectly correct: before IQs can make any sense as 'measures of intelligence', one would have to show *empirically* that there exists exactly one trait measured by all of them. Spearman was the first and also the last psychometrician to face up to this challenge. Thurstone only talked about it and his epigones apparently never even understood the problem. When Spearman failed, psychometrics had indeed lost its basic prop. At times Jensen (e.g., 1979) seems to have appreciated the

significance of Spearman's efforts (Jensen, 1979), if not the reasons for his failure. In this respect, and perhaps also in his untiring efforts to shore up his speculations with independent experimental evidence, Jensen is ahead of many of his peers.

As for his speculations about 'possible dysgenic trends' (Jensen, 1969, p. 91)—'have we seriously thought sufficiently of the rights of children . . . not to have a retarded parent . . . ?' (*ibid.*, p. 93)—we should perhaps thank Burt and Jensen for reminding us that arrogance, ignorance and prejudice have been fellow travellers of the mental testers ever since Galton (see, e.g., Chase, 1975), and that the frontiers of a responsible and rational social science remain yet to be conquered.

REFERENCES

- Boring, E. G. (1923) 'Intelligence as the tests test it', *New Republic*, **35**, pp. 35-7.
- Carroll, J. B. and Horn, J. L. (1981) 'On the scientific basis of ability testing', *American Psychologist*, **36**, pp. 1012-20.
- Chase, A. (1975) *The Legacy of Malthus: The Social Costs of the New Scientific Racism*, New York, Knopf: Illini Books edition, 1980, Urbana, Ill., University of Illinois Press.
- Guttman, L. (1977) 'What is not what in statistics', *The Statistician*, **26**, pp. 81-107; reprinted in Borg, I. (Ed.), (1981) *Multidimensional Data Representations: When and Why*, Ann Arbor, Mich., Mathesis, pp. 20-46.
- Jensen, A. R. (1966) 'Social class and perceptual learning', *Mental Hygiene*, **50**, 2, pp. 226-39.
- Jensen, A. R. (1969) 'How much can we boost IQ and scholastic achievement?', *Harvard Educational Review*, **39**, pp. 1-123.
- Jensen, A. R. (1979) 'g: Outmoded theory or unconquered frontier?', *Creative Science and Technology*, **11**, 3, pp. 16-29.
- Jensen, A. R. (1980) *Bias in Mental Testing*, New York, Free Press.
- Jensen, A. R. (1982a) 'The debunking of scientific fossils and straw persons', *Contemporary Education Review*, **1**, 2, pp. 121-35.
- Jensen, A. R. (1982b) 'Reaction time and inspection time measures of intelligence', Eysenck, H. J. (Ed.), *A Model for Intelligence*, New York, Springer.
- Jensen, A. R. (1983) 'The definition of intelligence and factor score indeterminacy', *The Behavior and Brain Sciences*, **6**, 2, pp. 313-15.
- Jensen, A. R. and Rohwer, W. D. Jr (1966) 'The Stroop Color-Word Test: A review', *Acta Psychologica*, **25**, pp. 36-93.
- Reichenbach, H. (1947) *Elements of Symbolic Logic*, New York, Macmillan; Dover Reprint, New York, Dover, 1975.
- Schönemann, P. H. (1981a) 'Power as a function of communality in factor analysis', *Bulletin of the Psychonomic Society*, **17**, 1, pp. 57-60.
- Schönemann, P. H. (1981b) 'Factorial definitions of intelligence: Dubious legacy of dogma in data analysis', in Borg, I. (Ed.), *Multidimensional Data Representations: When and Why*, Ann Arbor, Mich., Mathesis Press, pp. 325-63.
- Schönemann, P. H. (1983) 'Do IQ tests really measure intelligence?', *The Behavior and Brain Sciences*, **6**, 2, pp. 311-15.
- Schönemann, P. H. and Steiger, J. H. (1976) 'Regression component analysis', *British Journal of Mathematical and Statistical Psychology*, **29**, pp. 175-89.
- Schönemann, P. H. and Wang, M. M. (1972) 'Some new results of factor indeterminacy', *Psychometrika*, **37**, pp. 61-91.
- Spearman, C. (1927) *The Abilities of Man*, New York, Macmillan.
- Steiger, J. H. and Schönemann, P. H. (1976) 'A history of factor indeterminacy', in Shye, S. (Ed.), *Theory Construction and Data Analysis in the Social Sciences*, San Francisco, Calif., Jossey Bass, pp. 136-78.
- Thomson, G. H. (1916) 'A hierarchy without a general factor', *British Journal of Psychology*, **8**, pp. 271-81.
- Thorndike, E. L. (1903) *Educational Psychology*, New York, Teachers College, Columbia University.
- Tuddenham, R. D. (1962) 'The nature and measurement of intelligence', in Postman, L. (Ed.), *Psychology*

- in the Making: *Histories of Selected Research Problems*, New York, Knopf.
- Wechsler, D. (1939) *The Measurement of Adult Intelligence*, 1st ed., Baltimore, Md., Williams and Wilkens; reprinted in Edwards, A. J. (1974), *Selected Papers of David Wechsler*, New York, Academic Press.
- Wechsler, D. (1958) *The Measurement and Appraisal of Adult Intelligence*, 4th ed., Baltimore, Md., Williams and Wilkens.
- Wechsler, D. (1975) 'Intelligence defined and undefined: A relativistic appraisal', *American Psychologist*, **30**, pp. 135-9.
- Wilson, E. B. (1928) 'Review of *The Abilities of Man, Their Nature and Measurement* by C. Spearman', in *Science*, **67**, pp. 244-8; reprinted in Borg, I. (Ed.), (1981) *Multidimensional Data Representations: When and Why*, Ann Arbor, Mich., Mathesis Press.

ARTHUR JENSEN

Consensus and Controversy

EDITED BY

Sohan Modgil, Ph.D.

Reader in Educational Research and Development
Brighton Polytechnic

AND

Celia Modgil, Ph.D.


Senior Lecturer in Educational Psychology
London University

CONCLUDING CHAPTER

BY

Arthur R. Jensen

University of California, Berkeley

 The Falmer Press
(A Member of the Taylor & Francis Group)
New York Philadelphia and London