Multivariate Behavioral Research, 27 (2), 225-233

Spearman's Hypothesis: Methodology and Evidence

Arthur R. Jensen
University of California, Berkeley

Guttman's sophistic critique so badly misrepresents the case I have made for what I have called "Spearman's hypothesis" (Jensen, 1985a) that it is necessary here to spell out completely and accurately my position on this subject. Although I have already done this in much more detail in previous publications, Guttman seems not to have understood. I hope that readers will study the actual argument, methodology, and evidence for Spearman's (1927) hypothesis that I have presented in the several cited publications. Spearman's hypothesis is an empirical question testable by a clearly specified method applied to real psychometric data. Let me state my position on the main issues.

The General Factor

The most important factor in the cognitive domain is the general factor, or *g*. This is so for a number of reasons. Probably the most important is the fact that *g* is more highly correlated with various indices of learning, performance, and achievement outside the set of psychometric tests of mental abilities from which *g* is derived than is the case for any other factor or combination of factors (independent of *g*) that can be derived from the factor analysis of a given set of tests. In brief, *g* is the chief *active ingredient* in the concurrent and predictive validity of most psychometric tests in most of the situations in which tests are used. Also, the *g* factor accounts not only for a larger proportion of the common factor variance of various collections of diverse tests than any other factor, but often accounts for more of the common factor variance than all of the other factors combined. For example, in a study of 18 separate factor analyses of test batteries comprising anywhere from 6 to 13 tests, the *g* factor accounted on average for 4.3 times as much variance in test scores as all of the other common factors combined (Jensen, 1987a).

But it should also be noted that there is a great deal of uniqueness (i.e., specificity + error) in tests. In the study just mentioned, for example, tests' uniqueness accounts, on average, for nearly one half of the total variance in test scores. A test's specificity is usually problematic. It is often virtually impossible to characterize precisely in psychological terms. Moreover,

assuming a particular test was factor analyzed among a large and diverse battery of other tests, our knowledge of the particular test's specificity would probably have no value for most of the practical purposes for which tests are generally used. For most of the criteria ordinarily predicted by tests, a test's predictive validity would probably be reduced to nil if its general factor and major group factors were partialled out.

The predominance of *g* should not belittle the importance of other substantial group factors (e.g., verbal, spatial, memory) and special talents (e.g., musical, artistic, mechanical, motoric). However, it is a popular misconception that every person has such large peaks and valleys across the total spectrum of abilities that it is virtually impossible to speak realistically of different persons as being higher or lower in abilities in some average or general sense. But the very existence and size of the general factor absolutely contradicts this notion. It is a logical corollary of *g* that the average difference betweeen various abilities *within* individuals is smaller, in general, than the average difference *between* individuals in their overall average level of ability.

Now we must consider the three most commonly expressed doubts about the *g* factor.

Different Methods for Extracting g

In the modern psychometric literature *g* is represented by any one of three methodologically and conceptually rather different methods: (a) as the first principal component, (b) as the first principal factor (unrotated) in a common factor analysis (also called *principal factor* or *principal axes* analysis), and (c) as the second-order (or highest order) factor in an orthogonalized hierarchical factor analysis. It has been found to be true empirically (although it is not necessary mathematically) that the *g* extracted by any one of these methods is very highly correlated (usually .990 to .999) with the *g* extracted by either of the other methods in the same set of tests. It should be noted that Spearman's (1927) original but now outmoded method of factor analysis can be correctly applied only to the rare instances of reduced correlation matrices of unit rank. It is therefore useless for analyzing large and factorially complex matrices, and so of course I have never used Spearman's single factor method in testing what I have termed Spearman's hypothesis. Nor does this hypothesis in any way depend on Spearman's long defunct "two-factor" theory of mental abilities.

The empirical reality of nonzero positive correlations between all cognitive tests is the fundamental condition for a general factor, and any correlation matrix displaying this condition will yield a general factor by any one of these three contemporary methods of factor analysis just mentioned. (The only exceptions are those methods, like Thurstone's, 1947, multiple factor analysis

with orthogonal rotation of the factor axes, which necessarily submerge the general factor among the [orthogonally rotated] primary factors.) I have yet to find a bona fide empirical demonstration of negative correlations between cognitive ability tests that are significantly replicable or cannot be explained by some combination of sampling error and restriction of range on *g* in the subject sample.

Invariance of g

Although it is not a mathematical necessity, it is an empirical fact that the *g* factor is quite stable when extracted from different batteries of cognitive tests, provided the tests composing each battery are reasonably numerous and diverse in information content and task demands. In fact, the degree of invariance of *g* is a direct function of both the number and diversity of the tests. Also, a hierarchical *g* is generally somewhat more stable than either the first principal component or the first principal factor. I have found, for example, that estimated *g* factor scores derived from a factor analysis of just the six Verbal subtests of the Wechsler Adult Intelligence Scale (WAIS) are correlated .80 with the estimated *g* factor scores dervived from a factor analysis of just the six nonverbal Performance tests. Yet there is no resemblance between the Verbal and Performance subtests in their information content or specific task demands.

A large-scale investigation of *g* invariance was conducted by the late R. L. Thorndike (1987). He began with 65 highly diverse tests used by the U.S. Airforce. From 48 of these tests, six non-overlapping batteries were formed, each composed of eight randomly selected tests. Into each of these six batteries was inserted, one at a time, each of the 17 remaining *probe* tests. Hence each of the six batteries was factor analyzed 17 times, each time containing a different one of the 17 probe tests. The six *g* loadings obtained for each of the 17 probe tests then were compared with one another. It was found that the six *g* loadings for any given test were highly similar, although the *g* loadings varied considerably from one test to another. The average correlation between *g* loadings across the six batteries was .85. If each battery had contained more tests from the same general test pool, it is a statistical certainty that the average cross-battery correletions between *g* loadings would be still higher. Thorndike's finding, which is consistent with similar studies, constitutes strong evidence that pretty much the same *g* emerges from most collections of diverse cognitive tests. This evidence also indicates that the invariance of *g* across test batteries does not depend on their having *identical elements* in common, in the sense of elements of test content. Even highly dissimilar tests (e.g., vocabulary and

block designs) can have comparably high loadings on one and the same *g* factor.

Just as we can think statistically in terms of sampling error of a statistic when we randomly select a limited group of subjects from a population, or of measurement error when we obtain a limited number of measurements of a particular variable, so too we can think in terms of *psychometric sampling error*. In making up any collection of cognitive tests, we do not have a perfectly representative sample of the entire population of cognitive tests or of all possible cognitive tests, and so any one limited sample of tests will not yield exactly the same *g* as another limited sample. The sample values of *g* are affected by subject sampling error, measurement error, and psychometric sampling error. But the fact that *g* is very substantially correlated across different test batteries means that the variable values of *g* can all be interpreted as estimates of some true (but unknown) *g*, in the same sense that, in classical test theory, an obtained score is viewed as an estimate of a true score.

Is g an Artifact?

This question implies that *g* may have no significance or substantive meaning other than the mathematical technique used in deriving it. This is a false implication, for three main reasons.

First, a hierarchical general factor is not at all a mathematical necessity, and correlation matrices outside the cognitive realm can be found which yield no general factor. Therefore the presence (or absence) of a hierarchical *g* is itself an empirical fact rather than a trivial tautology. It simply reflects the all-positive correlations among tests in the matrix, a condition which is not forced by any methodological machinations.

Second, a highly replicable mathematical dimension that can be explicitly defined and empirically demonstrated under specified conditions is real. It is real in the same sense that other scientific constructs (e.g., gravitation, magnetic field, potential energy) are real and measurable, even though they are not directly observable or tangible entities.

Third, *g* is related to other variables and constructs which lie entirely outside the realm of psychometrics and factor analysis and have no connection whatsoever with these methodologies. For example, the degree to which various psychometric tests are *g* loaded is highly related to their degree of correlation with variables such as the *heritability* of individual differences in the test scores, the *spouse correlations* and various *genetic kinship correlations* in the test scores, the *effects of inbreeding* (and its counterpart, *heterosis*) on test performance, *choice reaction time* to visual and auditory stimuli, *inspection*

time (i.e., the speed of visual or auditory discrimination), and certain features of the brain's *evoked electrical potentials*. (These studies have been cited in Jensen, 1987b.) No other factor that can be extracted from a collection of diverse cognitive tests shows as large correlations with as many different biologic or other non-psychometric variables as does *g*. It is clear that *g* has as much claim to reality as theoretical constructs in other sciences. It is one of the major constructs in psychology, and one of the oldest and most well-established.

Spearman's Hypothesis

The *g* factor takes on further significance through its connection with an hypothesis first suggested by Charles Spearman (1927, p.379). Spearman noted that the average difference (in standardized score units) between representative samples of the black and white populations in the United States differ considerably from one test to another, and he commented that the size of these differences is directly related to the size of the *g* loadings of the tests on which the differences are found, regardless of the particular type or content of the tests.

I have formalized Spearman's (1927) original observation, calling it "Spearman's hypothesis" (Jensen, 1985a). It states that the relative magnitudes of the standardized mean black-white differences on a wide variety of cognitive tests are related predominantly to the relative magnitudes of the tests' *g* loadings — the higher the test's *g* loading, the larger the mean black-white difference. This hypothesis, if true, would mean that understanding the nature of the statistical black-white differences on various psychometric tests in the cognitive domain depends fundamentally on understanding the nature of *g* itself.

Methodology

A proper test of Spearman's (1927) hypothesis requires the following conditions:

1. The black and white samples must be fairly representative of their respective populations and should be sufficiently large that there is small enough sampling error of the correlations among tests to yield stable factors; and the samples should not be selected on any variables, such as educational or occuptional level, that would restrict the range-of-talent with respect to *g*.

2. The collection of psychometric tests should be fairly numerous, to permit the extraction of a relatively reliable *g* factor.

3. The various mental tests should be fairly diverse in content and task demands, both to insure a stable *g* and to allow considerable reliable variation in the *g* loadings of the various tests.

4. The tests' *g* loadings and the standardized mean group differences should be corrected for attenuation.

5. The factor analysis must be carried out *within* either the white or the black sample (or separately in both) but not in the combined samples, so that there is no possibility that any between-samples variance can enter into the correlations or the factor analysis of them.

6. The similarity in the vector of *g* loadings extracted separately from the two groups must be sufficiently high to assure that the same factor is represented in both groups, as indicated by a coefficient of congruence greater than +.90.

The statistical test of Spearman's hypothesis (Jensen, 1985a), then, is the rank order correlation between the tests' *g* loadings (in either group) and the standardized mean differences between the groups on each of the tests (with loadings and differences corrected for attenuation).

Empirical Evidence

I have investigated Spearman's hypothesis in eleven large data sets that meet these requirements, some more ideally than others (Jensen, 1985a, 1985b). The hypothesis was borne out in every study, and the results are approximately the same whether *g* is represented by the first principal component, the unrotated first principal factor, or the second-order hierarchical factor obtained from a Schmid-Leiman (Schmid & Leiman, 1957) orthogonalized hieracrchical factor analysis. The larger the number of tests and the greater the dispersion of the tests' *g* loadings, the more strongly the results accord with Spearman's hypothesis, that is, a large and significant positive correlation between (a) various tests' *g* loadings, and (b) the sizes of the tests' standardized mean differences between the white and black samples. Probably the most ideal set of data for testing Spearman's hypothesis consisted of samples of 4th and 5th grade black and white pupils who were matched on age, sex, school, and socioeconomic status (Naglieri & Jensen, 1987). A Schmid-Leiman orthogonalized hierarchical *g* was extracted from 24 diverse mental tests within each racial group. As the congruence coefficient between the black and white groups was .95, the *g* loadings of each test were averaged across racial groups. This vector of 24 *g* loadings had a Pearson *r* of +.78 and a Spearman rank-order correlation of +.75 with the sizes of the 24 standardized mean differences between the white and black groups on the tests. Despite assiduous search, no set of data appropriate for testing Spearman's hypothesis

has yet been found that fails to support the hypothesis. Hence the hypothesis is now so strongly confirmed as to be regarded as an empirical fact.

My formalization (or reformulation) of Spearman's hypothesis (Jensen, 1985a), it is important to note, states that the variation in the mean black-white differences on various tests is associated *predominantly* (rather than exclusively) with the tests' *g* loadings. This weaker version of the hypothesis is dictated by the empirical finding that when we plot the linear regression of black-white differences on tests' *g* loadings, we find that certain tests consistently show deviations from the regression line. Tests that have an appreciable loading on a *spatial* factor (e.g., block designs, object assembly, paper folding, comparison of rotated figures, and the like) consistently show a *larger* black-white difference than is predicted from the test's *g* loading. Tests with an appreciable loading on a *short term memory factor* (e.g., digit span, verbal rote learning, digit symbol or coding) show a *smaller* black-white difference than is predicted by the test's *g* loadings. So far, these are the only two well-established psychometric factors that have been found to cause rather small but consistent perturbations in demonstrations of Spearman's hypothesis.

Additional analyses (Jensen, 1987a) further substantiate Spearman's hypothesis. Into each of 18 independent correlation matrices, comprising anywhere from 6 to 13 tests (averaging 11.1 tests), with each matrix based exclusively on either a white or a black sample (but never a racially mixed sample), were inserted the point-biserial correlations of each of the tests in the particular matrix with the variable of race treated as a dichotomous variable (quantitized as black = 0, white = 1). Each matrix was subjected to a principal factor analysis with a minimum of four first-order factors extracted from each matrix. The average loading of the dichotomous race variable on the *g* factor was .55, whereas the average of the corresponding loadings on the three largest first-order factors (all uncorrelated with *g*) was .24. In other words, the black/white variable generally had its major loading on the *g* factor. A spatial visualization factor is the only non-*g* factor that rather consistently rivals *g* in its loadings on the black/white variable (see also Naglieri & Jensen, 1987). Hence the largest black-white mean difference is seen on those tests that are the most highly loaded on both *g* and a spatial factor. The smallest black-white mean differences occur on tests that are the least loaded on *g* and the most highly loaded on a short-term memory factor. Contrary to popular belief, the mean black-white difference on the verbal factor (independent of *g*) is nil. Examination of 121 psychometric tests that were factor analyzed in eleven studies also showed that the *g* loadings of various tests are distributed as a continuous variable extending over a wide range of values — from about .30 up to nearly .90. On the same set of tests, the black-white mean differences (expressed in standard deviation units) are also distributed as a continuous

variable, ranging from close to zero up to about 1.3 standard deviations (*SD*s). From the linear regression of the mean black-white differences on tests' *g* loadings, the estimated mean difference on a hypothetical pure measure of *g* would be approximately 1.2 *SD*s.

Thirty-six scholars have published *peer reviews* of my research on Spearman's hypothesis (1985a, 1987a), but none has refuted the methodology or the empirical demonstration of Spearman's hypothesis.

Guttman's Critique

Scarcely anything that Guttman has to say about my treatment of Spearman's hypothesis (Jensen, 1985a) is relevant to what I have actually done or written with respect to it. As anyone who reads the cited articles will readily realize, Guttman's peevish critique is directed at a straw man — a muddled and misleading misrepresentation of my method of testing Spearman's hypothesis. Chiefly, he promotes the misapprehension that the demonstration of the hypothesis is somehow a mathematical necessity or tautology rather than an empirical discovery, and that confirmation of the hypothesis was an inevitable result of the methodology for testing the hypothesis and hence is purely an artifact. This claim is clearly impossible, for two quite obvious reasons:

1. When Pearson correlation coefficients between tests are calculated, all information about the means and standard deviations of groups' means on the tests is completely lost. The same logic, of course, necessarily applies to the differences between groups' means on the various tests. Consequently, nothing about the groups' means, or the differences between groups' means on the tests or the rank order of their magnitudes, can be inferred from the matrix of test intercorrelations. Ipso facto, nothing can be mathematically inferred about the rank order of tests' means (or mean group differences) from a knowledge of the tests' loadings on *g* or on any other factors extracted from the correlation matrix.

2. The test means of one or the other comparison group (either black or white) are also *experimentally independent* of the data from the group which yielded the test intercorrelations and the *g* factor extracted from them.

These two self-evident statistical facts necessarily mean that the prescribed method for testing Spearman's hypothesis (Jensen, 1985a) yields a result that cannot be an artifact or a tautology. Hence a reliable correlation between tests' *g* loadings and the standardized mean differences between groups on these tests is necessarily a genuine phenomenon. Since Spearman's hypothesis has been consistently borne out in many independent sets of appropriate data, and no contrary data have been found, it may legitimately claim the status of empirical fact.

References

Jensen, A. R. (1985a). The nature of the black-white difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences, 8,* 193-219.

Jensen, A. R. (1985b). The black-white difference in *g*: A phenomenon in search of a theory. *Brain and Behavioral Sciences, 8,* 246-263.

Jensen, A. R. (1987a). Further evidence for Spearman's hypothesis concerning black-white differences on psychometric tests. *Behavioral and Brain Sciences, 10,* 512-519.

Jensen, A. R. (1987b). The *g* beyond factor analysis. In J. C. Conoley, J. A. Glover, & R. R. Ronning (Eds.), *The influence of cognitive psychology on testing and measurement.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Naglieri, J. A., & Jensen, A. R. (1987). Comparison of black-white differences on the WISC-R and the K-ABC: Spearman's hypothesis. *Intelligence, 11,* 21-43.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 53-61.

Spearman, C. (1927). *The abilities of man.* London: Macmillan.

Thorndike, R. L. (1987). Stability of factor loadings. *Personality and Individual Differences, 8,* 585-586.

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago, IL: University of Chicago Press.