Commentary

Humphreys's Attenuated Test of Spearman's Hypothesis

ARTHUR R. JENSEN

School of Education
University of California, Berkeley
Berkeley, CA 94720

Humphreys's test of Spearman's hypothesis (viz., that the size of the standardized black–white difference on various psychometric tests is positively related to the tests' loadings on g, the general intelligence factor) is methodologically weak. It is based on comparison of a fairly representative sample of the black population of U.S. school children with a highly selected sample of the white school population, representing the lowest 15 to 20% of whites in socio-economic status (SES). A fair test of the hypothesis requires that the black and white samples should not be selected on any g-correlated variable, such as SES. Selection attenuates and distorts the relationship between tests' g loadings and the magnitude of the standardized mean black-white differences on the tests. Other unorthodox conditions in Humphreys's study, such as performing factor analysis on the test-score means of various arbitrary demographic groups instead of on individual test scores inflates tests' g loadings and biases the test of Spearman's hypothesis by restriction of reliable variance in g loadings. Humphrey's study cannot be considered a proper replication of Jensen's examination of Spearman's hypothesis in 11 different studies comprising 74 different tests, which consistently bears out Spearman's hypothesis.

Charles Spearman originally suggested in 1927 that the varying magnitude of the mean differences between whites and blacks in standardized scores on a variety of mental tests is directly related to the size of the tests' loadings on g, the general factor common to all complex tests of mental ability. I have examined this hypothesis in 11 sets of published data on black–white diffgrences on 121 tests (74 of them different tests) with samples totalling 29,712 white and 10,783 black Americans (Jensen, 1985). These studies consistently bear out Spearman's hypothesis of a positive correlation between various tests' g loadings and the size of the average black–white differences on the various tests. The evidence suggests that the black–white difference is more a difference in Spearman's g than a difference in narrower group factors associated with any particular item content, cultural knowledge, specific skills, or type of test (also see Jensen & Reynolds, 1982).

There are usually three legitimate and straightforward methods for testing

---

Spearman's hypothesis: (1) correlation of tests' g loadings with the standardized mean black–white differences (the rank–difference correlation affords a strict test of significance, as it is strictly a permutation test that does not depend on any assumptions about the distributional characteristics of the correlated variates); (2) the means of factor scores on g and on any other significant factors in the test battery are compared between black and white samples; (3) the point-biserial correlation (corrected for restriction due to unequal sample sizes) between the black/white classification and test scores are entered into the factor anlaysis of all of the tests, and the size of the loading of the black/white variable on the g factor is compared with its loading on other significant factors in the matrix. (In number 1, the point-biserial correlation between the black/white variable and test scores may be substituted for the standardized mean difference, since there is a nearly linear relationship between these two parameters in the range of mean test score differences typically found between black and white populations.)

Application of any one or a combination of these methods to the data of 11 studies described elsewhere (Jensen, 1985) indicates that g is the predominant factor in the black–white difference. These tests of the hypothesis, however, are of questionable legitimacy and cogency when they are not applied to reasonably representative samples of the black and white populations. In my first discussion of Spearman's hypothesis (Jensen, 1980, p. 535), I pointed out that "the Spearman hypothesis cannot be appropriately tested on samples that were in any way specially selected with reference to any g-loaded characteristics." For example, we could not properly test the hypothesis on black and white students in a selective college that applies the same selection criteria to all applicants, since such selection for academic aptitude would tend to equalize the population means on the most g-loaded tests. Any selection of samples on general ability would work directly against Spearman's hypothesis to some degree, and the g factor extracted from tests given to selected groups would be considerably diminished and probably distorted, as compared with the g extracted from the same tests given to random samples of either the black or the white population.

Humphreys's (1985) study violates this condition by comparing a fairly representative sample of blacks with a sample of very low socio-economic status (SES) whites, representing the lowest 15 to 20% of the total white sample of the Project TALENT data bank. The least we should expect of a replication of previous tests of Spearman's hypothesis is that the same methods be used, in addition to any new analyses that may be proposed.

We know from much other evidence that low-SES whites more nearly resemble blacks in general mental ability, or g, than does the total white population. The g factor is correlated with SES, and the deck is stacked against Spearman's hypothesis by comparing blacks with only low-SES whites. But other methodological differences have also probably biased Humphreys's results against the hypothesis. Two of the eleven studies (Hennessy & Merrifield, 1976; Reynolds

& Gutkin, 1981) analyzed in Jensen's (1985) study either matched or statistically equated black and white samples on SES and yet showed much higher rank-order correlations ($+.59$ and $+.54$, respectively) between g loadings and mean black–white differences than the correlation of $+.17$ found in Humphreys's study. What really needs to be explained, then, is the reason for Humphreys's anomalous or attenuated correlation, which is an outlier, well out of bounds when compared with eleven other large data sets, including even those less-than-ideal data that equated blacks and whites on SES.

Although it is not a part of Spearman's hypothesis, it is not at all surprising that white upper-SES and lower-SES groups also differ directly in proportion to the tests' g loadings. This striking finding bears out the theory that g is an important factor in social mobility via educational attainment and occupational status. Comparing the total black sample with a low-SES white sample naturally makes their test profiles more alike, especially diminishing the differences on the most g-loaded tests. We can see this quite clearly by looking at the profiles of white and black means on the Wechsler Intelligence Scale for Children-Revised (WISC-R) in the total national standardization data. Here, the correlation between the white and black profiles is $-.76$. If one obtains a subsample of whites who are matched with the total black sample on IQ, however, the correlation between the profiles of WISC-R subtest means is $+.53$. Matching blacks and whites on SES, of course, would be less drastic, but would move the correlation between the black and white profiles closer toward zero. For example, the midpoint between these extreme correlations (i.e., $-.76$ and $+.53$) is $-.23$, which is close to the correlation of $-.19$ between the black versus low-SES white profile correlation reported by Humphreys. (The WISC-R profiles referred to can be found in Jensen and Reynolds [1982] and Reynolds and Jensen [1983].) In Humphreys's Project TALENT data, a comparison of the black sample with the upper-SES white groups, representing the upper 80 to 85% in SES, would have provided a somewhat more suitable, though still less than ideal, test of the hypothesis. Unfortunately, the correlation between g loadings and the mean differences between blacks and high-SES whites was not reported. In fact, the results in general are inadequately reported; the least that should have been provided is a table giving the names of the tests in the Project TALENT battery, the tests' g loadings in the black and white samples, and the standardized mean black–white difference on each test (or the point-biserial correlation of each test with the black/white classification).

Another methodological condition that has generally been observed in previous tests of the hypothesis is that the g factor has been extracted in both the black and white samples separately and the hypothesis is tested only if the g factor is found to be essentially the same in both groups, as shown by a high coefficient of congruence. Humphreys presents no such demonstration in his study. It is also important to extract g separately in the groups being compared so

that the g-factor loadings are not contaminated with variance due to the average difference between groups. There is no indication that Humphreys observed this caution. Also, no account is taken of differences in test reliability which can differentially attenuate both g loadings and standardized mean differences.

Humphreys's data consist entirely of the mean test scores obtained in numerous schools. The data have not been analyzed at the level of individual differences. A g factor extracted from correlations between means, so-called "ecological correlations" (Robinson, 1950), is bound to have markedly inflated g loadings, and, since there is a ceiling for factor loadings, the loadings are mostly compressed into the high end of the scale, restricting their variance and thereby attenuating their correlation with the black–white differences. A correlation scatter diagram of these data provided by Humphreys (personal communication) reveals that more than half of the g loadings are greater than $+0.75$, and with this restricted range of g loadings, some 70% of all the loadings appear to be ties in the scatter plot; thus, factor analysis of group means works against finding reliable differences between g loadings, and weakens the correlation between the g loadings and the black–white differences. Moreover, ecological correlations and factors derived from aggregated data are not necessarily equivalent to the correlations and factors derived from individual data; equivalence is implicitly assumed in Humphreys's analysis.

The suitability of some of the tests in the Project TALENT battery may also be questioned as vehicles for testing Spearman's hypothesis. Some of these tests are very short, relatively unreliable, and are designed to assess such narrow and highly culture-loaded content as knowledge about domestic science, farming, fishing, hunting, and mechanics. A serious psychometric deficiency of some of these tests is a "floor effect" in the black sample, that is, test items are too difficult to allow measurement of the full range of ability in the black sample, a condition that has the effect of spuriously diminishing the observed difference between the black and white means on certain tests. A scale transformation of these data cannot really remedy the undesirable "floor effect" in the black sample.

Although Spearman's hypothesis per se, does not concern the *cause* of the black–white difference in g, and does not make any statement about SES differences within racial groups, it is interesting, from the standpoint of the theory that g is strongly involved in social mobility (via occupational status), that a high-positive correlation is found between tests' g loadings and the differences between the high-SES and low-SES subgroups of the white sample. With respect to Spearman's hypothesis per se, however, Humphreys's treatment seems scarcely adequate.

A more thorough theoretical and methodological discussion of Spearman's hypothesis, along with analyses of data from 11 studies bearing on it, is presented in my article (Jensen, 1985) in the June, 1985 issue of *The Behavioral and*

Brain Sciences, which includes critical commentaries on my article by 27 behavioral scientists. I will gladly send a reprint of this article with the commentaries, as well as reprints of my other related articles cited here, to anyone who requests them.

REFERENCES

Hennessy, J. J., & Merrifield,. P. R. (1976). A comparison of the factor structures of mental abilities in four ethnic groups. *Journal of Educational Psychology, 68,* 754–759.

Humphreys, L. G. (1985). Race differences and the Spearman hypothesis. *Intelligence, 9,* 275–283.

Jensen, A. R. (1980). *Bias in mental testing.* New York: The Free Press.

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

Jensen, A. R., & Reynolds, C. R. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences, 3,* 423–438.

Reynolds, C. R., & Gutkin, T. B. (1981). Multivariate comparison of the intellectual performance of blacks and whites matched on four demographic variables. *Personality and Individual Differences, 2,* 175–180

Reynolds, C. R., & Jensen, A. R. (1983). WISC-R subscale patterns of abilities of blacks and whites matched on full scall IQ. *Journal of Educational Psychology, 75,* 207–214.

Robinson, W. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15,* 351–357.