THE BEHAVIORAL AND BRAIN SCIENCES (1985) 8, 193–263
*Printed in the United States of America*

# The nature of the black–white difference on various psychometric tests: Spearman's hypothesis

**Arthur R. Jensen**
*School of Education, University of California, Berkeley, Calif. 94720*

**Abstract:** Although the black and white populations in the United States differ, on average, by about one standard deviation (equivalent to 15 IQ points) on current IQ tests, they differ by various amounts on different tests. The present study examines the nature of the highly variable black–white difference across diverse tests and indicates the major systematic source of this between-population variation, namely, Spearman's *g*. Charles Spearman originally suggested in 1927 that the varying magnitude of the mean difference between black and white populations on a variety of mental tests is directly related to the size of the test's loading on *g*, the general factor common to all complex tests of mental ability. Eleven large-scale studies, each comprising anywhere from 6 to 13 diverse tests, show a significant and substantial correlation between tests' *g* loadings and the mean black–white difference (expressed in standard score units) on the various tests. Hence, in accord with Spearman's hypothesis, the average black–white difference on diverse mental tests may be interpreted as chiefly a difference in *g*, rather than as a difference in the more specific sources of test score variance associated with any particular informational content, scholastic knowledge, specific acquired skill, or type of test. The results of recent chronometric studies of relatively simple cognitive tasks suggest that the *g* factor is related, at least in part, to the speed and efficiency of certain basic information-processing capacities. The consistent relationship of these processing variables to *g* and to Spearman's hypothesis suggests the hypothesis that the differences between black and white populations in the rate of information processing may account for a part of the average black–white difference on standard IQ tests and their educational and occupational correlates.

**Keywords:** black–white differences; factor analysis; individual differences; information processing; mental chronometry; intelligence; psychometric *g*; psychophysiology; reaction time

In representative samples of native-born black and white Americans, the latter have scores that are an average of about one standard deviation higher than the former in the distribution of scores on standard psychometric tests of general intelligence and tests of scholastic aptitude and achievement (Jensen 1973a, Chap. 7; Loehlin, Lindzey & Spuhler 1975; Osborne & McGurk 1982; Shuey 1966). One standard deviation (*SD*) difference between the means of two approximately normal distributions with approximately equal *SD*s corresponds to a median overlap of about 16%, that is, 16% of the scores in the lower distribution surpass the median score (or 50th percentile) of the higher distribution.

Not all psychometric tests of ability, however, show the same mean difference (in *SD* or $\sigma$ units) or the same median overlap between the black and white populations, even when the very same samples are compared on various tests.[1] There is significant variation in the magnitude of the black–white difference from one test to another (Loehlin et al. 1975). For example, the population means differ in varying degrees on tests of various abilities, such as the Verbal, Reasoning, Spatial, and Numerical subscales of the Differential Aptitude Tests (Lesser, Fifer & Clark 1965). The average black–white difference is also much smaller on what I have termed Level I abilities (short-term memory and rote learning) than on Level II abilities (reasoning, abstraction, and problem solving) (Jensen 1973b; 1974a).

Differential psychologists have made little systematic effort to understand these statistically significant and fairly consistent variations among tests with regard to the size of the mean black–white difference. For many years now the most popular explanations have invoked cultural and linguistic differences. The tests showing the largest group differences are claimed to be biased against many black persons because of their emphasis on white middle-class cultural content, and the standard English used in verbal tests is claimed to be a less familiar and less appropriate testing medium for black testees. The explanatory power of these two hypotheses, however, has failed when the predictions that should logically follow from them have been empirically tested. When items from standard tests have been classified or rated by expert judges in terms of the items' cultural content, independently of any knowledge of the actual item statistics, the ratings of items' cultural loadings are not positively related to the items' black–white discriminability (Jensen 1977a; Sandoval & Miille 1980). In fact, McGurk (1951; 1953a; 1953b) found just the opposite. McGurk asked a panel of 78 judges, including professors of psychology and sociology, to classify 226 items from several well-known standardized tests of general intelligence into three cate-

 0140-525X/85/020193-71/$06.00 

gories: least cultural, neutral, and most cultural. (The meaning of "cultural" was left up to the subjective judgment of the raters.) The 184 items on which there was highest agreement among the judges as to the items' being most or least cultural were administered as a test to large samples of black and white high school pupils. It turned out that the mean black–white difference on the test composed of items classified as "least cultural" was almost twice as great as the mean black–white difference on the test composed of items classified as "most cultural." Obviously, there must be some property on which these two classes of items differ, apart from their judged cultural loading, that would account for this surprising result.

The claim that the style of language used in most standard verbal tests contributes to the population difference should lead to the expectation of a greater black–white difference on verbal than on nonverbal tests. However, the massive evidence on this issue is unequivocally counter to this expectation. McGurk (1975) has reviewed virtually the entire published literature between 1951 and 1970 regarding the median overlap between black and white score distributions on verbal and nonverbal intelligence tests. Surprisingly, the percentage overlap is *greater* for verbal (19%) than for nonverbal tests (15%) – a difference significant beyond the .01 level. Thus, in actual fact, the black–white differential is slightly smaller on verbal than on nonverbal tests. However, Jensen (1974b) found that when items from what is generally deemed a highly culture-loaded verbal test (the Peabody Picture Vocabulary Test) and items from a relatively culture-reduced nonverbal test (Raven's Colored Progressive Matrices) were perfectly matched for difficulty in a white sample of elementary school children, and then administered to a sample of black children in the same grades, there was no significant difference between the mean scores of the black pupils on the verbal and nonverbal tests. In other words, verbal and nonverbal tests that were perfectly matched in difficulty at the item level for white testees were thereby also matched in difficulty for black testees. Obviously, the mean black–white difference in the test scores is not closely linked to the verbal–nonverbal dimension of test characteristics. There has been the same sort of findings with the Wechsler Intelligence Scale for Children–Revised (WISC-R). When large black and white samples were either statistically equated or actually matched on Full Scale IQ, they showed no significant differences on such highly verbal subtests as Information, Similarities, and Vocabulary (Jensen & Reynolds 1982; Reynolds & Jensen 1983).

If variation in the mean black–white difference on various tests cannot be attributed either to variation in the tests' cultural loading or to the tests' degree of dependence on language, then we must inquire which other characteristic of tests or test items is primarily responsible for the population difference.

## Spearman's hypothesis

Charles Spearman (1863–1945) was one of the most creative intellects in the history of psychometrics. He gave us factor analysis, the rank-order correlation coefficient, the correction for attenuation, and the precise formulation of the relationship between the length of a test and its reliability. With his formulation of the "noegenetic laws" of cognition he can also be credited as a leading pioneer in what we today term cognitive theory (Spearman 1923). It also happens that he expressed an interesting and potentially unifying insight into the nature of the variation in the size of the black–white difference on diverse mental tests. Considering Spearman's excellent track record in psychometrics, it might pay to take another look at his original conjecture on the subject of black–white differences.

Spearman, in 1927, suggested a hypothesis concerning the nature of the black–white difference, which, as far as I can determine, has never been subjected to empirical investigation beyond Spearman's original observation. In commenting on a study by Pressey and Teter (1919), in which 10 diverse mental tests were administered to large samples of black and white American children, Spearman noticed that the black children, on average, obtained lower scores than the white children on all 10 tests. But he also noticed that the mean difference "was most marked in just those [tests] which are known to be most saturated with *g*" (Spearman 1927, p. 379). The smallest difference was on a test of rote memory, the largest on a test of verbal ingenuity (Disarranged Sentences). The first test had been found to be the poorest of the 10 tests in differentiating mentally retarded and average persons, whereas the second test was the most discriminating. Since Spearman's observation was based on a rather limited and unreplicated set of data, it seems best to regard it not as an empirical generalization but as a hypothesis. I shall henceforth refer to it simply as Spearman's hypothesis.

## The nature of *g*

The *g* factor is Spearman's label for the single largest independent source of individual differences that is common to all mental tests, regardless of form, content, or sensorimotor modality. It is the *general* (hence *g*) factor in any collection of tests, whether their items consist of verbal, numerical, spatial, pictorial, or any other content, provided they require some minimal degree of mental effort and there is an objective criterion of superior performance. Few present-day psychometricians would disagree with the conclusion expressed by Sternberg and Gardner (1982): "We interpret the preponderance of the evidence as overwhelmingly supporting the existence of some kind of general factor in human intelligence. Indeed, we are unable to find any convincing evidence at all that militates against this view" (p. 231). Because the *g* factor is, in a sense, a distillate of the variance that is common to any large collection of diverse tests, it tends to minimize sources of variance attributable to specific prior learned content, skills, talents, or interests. Most of the variance associated with these features turns up, in the factor analysis model, in the so-called group factors or in the specificities.

Spearman's *g* is surely one of the most interesting and enduring constructs in all of psychology. Unfortunately, our present knowledge about the nature of *g* is limited to descriptions of the types of tests or problems that are most *g*-loaded and to the general characteristics of the cog-

nitive processes that most highly *g*-loaded tasks seem to have in common. Spearman (1927) described these characteristics as "the eduction of relations and correlates" (i.e., inductive and deductive reasoning) and "abstractness." But I should emphasize here that Spearman's so-called two-factor theory of mental ability and his theory of *g* as a kind of general "mental energy" are of no particular relevance or importance in the present discussion. At this point, *g* need not be attributed any meaning beyond its operational definition in terms of factor analysis. The nature of *g* in terms that are independent of factor analysis is a separate theoretical issue subject to empirical study in its own right. Little, if anything, is as yet known about the physiological and biochemical substrate of *g*, although some empirically testable theories have been proposed (e.g., Eysenck 1982a). What we do already know about *g* with some assurance, however, is that its measurement does not depend on any particular test or on types of test or on any particular item contents. These all are merely vehicles, and *g* can be measured by a virtually unlimited variety of vehicles. Nor does the elicitation of *g* depend on specific acquired knowledge or skills. As a psychological construct, *g* cannot be adequately defined in terms of specific types of information, items of knowledge, specialized skills, or particular cognitive strategies. As David Wechsler (1958) has remarked, "Unlike all other factors [g] cannot be associated with any unique or single ability; *g* is involved in many different types of ability; it is in essence not an ability at all, but a property of the mind" (p. 124). While not yet well understood theoretically, *g* is unquestionably the single largest source of individual differences in all cognitive activities that involve some degree of mental complexity and that eventuate in behavior which can be measured in terms of some objective standard of performance.

Although a great deal more could be said about *g*, a few of the most salient findings, which I have documented elsewhere (Jensen 1980a, Chaps. 6 and 8), will be presented here, as background for the present study.

• The fundamental observation giving rise to *g* is the positive manifold phenomenon; that is, the existence of positive correlations between all tests in the cognitive domain, over a wide range of diversity, regardless of the content or other surface characteristics of the tests themselves. The *g* factor represents this salient fact of nature better than any other single factor or any combination of multiple orthogonal factors (which disperse the *g* variance among a number of primary factors and thus artificially create the misleading impression that there are zero correlations among the several clusters of tests defining the primary abilities).

• Taken together, the *g* factor *plus* smaller group factors (primary abilities independent of *g*) best represent the fact that, on average, overall differences *between* individuals in the population are greater than the differences among various abilities *within* individuals. Multiple orthogonal (i.e., uncorrelated) factors, without *g*, would not lead us to this (empirically established) expectation.

• Certain tests (generally those involving greater complexity of mental manipulation) are consistently more *g*-loaded than others when factor-analyzed in different batteries of various tests. Cognitively less complex tests (usually involving sensorimotor skills or rote learning ability) have rather consistently weak *g* loadings.

• Essentially the same *g* emerges from collections of tests that are superficially quite different, such as the Verbal and Performance subtests of the Wechsler Intelligence Scales, for which the *g* factor scores are correlated about +0.80. Unlike *all* other factors, *g* is not tied to any particular type of item content or acquired cognitive skill. (This is the basis for Spearman's theorem of "the indifference of the indicator" of *g*.)

• It has proved impossible to construct a test to measure any of the primary mental abilities (or first-order factors) that does not also measure *g* (Eysenck 1939). That is to say, scores on so-called factor-pure tests (i.e., tests designed to measure some single factor other than g) always measure *g* in addition to whatever primary ability factor they were specifically devised to measure. In tests of the primary mental abilities, moreover, the *g* variance is generally greater than the variance attributable to the primaries per se (e.g., verbal, numerical, spatial, memory). However, it has proved possible to devise tests that measure *g* and little or nothing else.

• The *g* factor reflects more of the variance observed in informal, commonsense estimates of intelligence, by parents, teachers, employers, and peers, than any other factor that can be extracted from psychometric tests. There is considerable commonality between psychologists' technical conceptualization of intelligence and the meanings attributed to "intelligence" by laymen (Sternberg et al. 1981). In addition, *g* discriminates more accurately than any other factor between average persons and persons diagnosed as mentally retarded by independent, nontest criteria, and between average persons and those who are recognized as intellectually gifted on the basis of their accomplishments.

• There is no general factor of human learning ability that is different from, or independent of, the *g* of psychometric tests. However, there is much more "specificity" (i.e., variance not related to any common factors) in various laboratory learning tasks than in most psychometric tests composed of numerous items.

• Although *g* may not be equally valued in all cultures, individual differences in *g*-related abilities can be recognized even by persons in societies that differ widely from Western industrial civilization (Reuning 1972).

• In its practical ability to forecast the success of individuals in school and college, in armed forces training programs, and in employment in business and industry, *g* carries far more predictive weight than any other factor or any other combination of factors independent of *g* (Jensen 1984a). This means that many "real-life" kinds of performance, and not just psychometric tests, are substantially *g*-loaded.

• As Humphreys (1981; 1983) has pointed out, even where mental tests are not implicated, the naturally occurring educational and occupational selection in our society involves *g* more than any other measurable psychological variable. Each "sieve" in educational and oc-

cupational screening selects on g, and this observation is as applicable in communist countries where mental ability tests are officially forbidden as it is in the United States. For this and other reasons, Humphreys aptly refers to g as "*the* primary mental ability."

• The genetic phenomenon of inbreeding depression (i.e., the diminution of a metric character in the offspring of genetically related parents, such as siblings or cousins) is indicative of genetic dominance of the genes enhancing the trait in question and suggests that during the course of human evolution there has been directional selection for genes that enhance the trait. Large-scale data on the offspring of cousin matings show that the degree of inbreeding depression observed on 11 diverse subtests of the Wechsler Intelligence Scale for Children is positively and significantly correlated with the subtests' g loadings (Jensen 1983a). (This is equally true whether g is extracted as a first principal factor or as a hierarchical second-order factor.)

• The g factor (and g factor scores) are substantially correlated with measures of the speed of information processing in simple laboratory tasks, such as simple and choice reaction times, which bear no resemblance to the usual psychometric tests from which the g is extracted (Carlson & Jensen 1982; Jensen 1979; 1980b; 1981; 1982a; 1982b; Jensen & Munro 1979; Nettelbeck & Kirby 1983; Vernon 1981b; 1983). It has been found, in a sample of 100 university students, that speed of information processing, as measured by reaction-time techniques, is correlated about 0.5 (or 0.7 when corrected for restriction of range) with the g factor of the Wechsler Adult Intelligence Scale (WAIS) and that no additional component of variance in the 12 WAIS subtests (including the verbal, performance, and memory factors) shows a significant correlation with the reaction time measures (Vernon 1983). At an even more basic level, there is now considerable evidence that g is correlated with such physiological variables as the amplitude, latency, and complexity of averaged evoked potentials in the brain, as measured by means of EEG apparatus and electrodes attached to the scalp (e.g., Callaway 1975; Eysenck 1982a; Hendricksen & Hendricksen 1980; Jensen, Schafer & Crinella 1981; Schafer 1982; Shucard & Horn 1972).

The fact that the g factor, more than any other factor, is related to variables such as choice reaction time, the average evoked potential, and inbreeding depression – variables whose origin and measurement are entirely independent of factor analysis – suggests that g is not merely a theoretically empty mathematical artifact of factor analysis but a construct laden with theoretical significance that extends well beyond the algebraic operations involved in its extraction from the intercorrelations among psychometric variables (Jensen 1983b).

## Elementary cognitive tasks (ECTs) and *g*

The correlation of psychometric g with reaction time and other chronometric variables derived from elementary cognitive tasks (ECTs) suggests the existence in all cognitive tasks of a common mechanism that causes individual differences in performance to be positively intercorrelated and hence allows the emergence of a general factor. Individual differences in performance on the ECTs used in chronometric studies are attributable not mainly to the intellectual content of the ECTs but to the speed or efficiency with which the ECTs are performed. ECTs are extremely simple laboratory tasks that are specially devised to measure response latencies in making decisions reflecting such elementary cognitive processes as stimulus apprehension, stimulus encoding and transformation, short-term memory scanning, retrieval of highly overlearned words from long-term memory, discrimination, mapping of semantic or spatial relations, and the like (Jensen, in press). Our own laboratory tasks are so simple that response latencies for young adults are generally less than one second. Yet highly reliable individual differences in response latencies emerge when averaged over a number of trials. Individual variation in speed of response to ECTs extends far beyond the range of variation in response time that can be accounted for in terms of sensory lag, speed of neural conduction in sensory and motor pathways, and muscle latency. Thus individual differences in response speed to ECTs appear to be largely of central origin. This is true even of simple reaction time.

That the speed of cognitive processes is related to physiological processes at the interface between brain and behavior is suggested by the evidence of average evoked potentials and the effects of physiological variations on reaction times. At present, there are only highly speculative theories as to the nature of these physiological mechanisms – the theory of errors or "noise" in the transmission of neural impulses (Eysenck 1982a), for example, or the theory of neural oscillations (Jensen 1982b).

Theorization at the psychological level of information processing is far more highly developed, however. One central theory holds that the speed and efficiency with which persons can execute the various elementary cognitive processes called for by ECTs are correlated with performance on highly diverse g-loaded psychometric tests because successful performance on all such tests, however markedly they may differ in appearance and surface content, depends on the execution of a number of shared or common underlying cognitive processes.

A crucial construct in this theory, which attempts to explain the correlation between mental speed, as measured in ECTs, and scores on complex psychometric tests of intelligence, is what has been termed "working memory" in theories of information processing. Working memory is understood to be a short-term memory system with a distinctly limited capacity for processing incoming information or information retrieved from long-term memory. Without continuous rehearsal, the limited information in working memory rapidly decays beyond retrieval and must be replaced by new input. Not only does the process of mentally manipulating the information being held in working memory absorb some of its capacity for processing incoming information, but every mental operation takes up a certain amount of time, and if common processes are involved in two or more different operations, these must be performed successively to avoid interference with successful execution of the operations. Overloading the capacity of the system causes shunting or inhibition of the information input or a momentary break-

down in internal operations. These effects have been demonstrated experimentally in many studies and are quite generally acknowledged as established phenomena in experimental cognitive psychology (Posner 1966; 1978; 1982).

How then do these limitations of working memory figure in the observed correlation between mental speed in various ECTs and performance on untimed psychometric tests? A faster rate of mental processing (e.g., encoding stimuli, chunking, transformation, and storage of incoming information and retrieval of information from long-term memory [LTM]) would presumably permit the system to compensate, in effect, for its limited capacity, by allowing critical operations to occur *before* the decay of information in working memory. At a slower rate of processing, the trace would decay before the solution was achieved, and repetition of the information input would be required until the correct response could occur. Memory span for recalling digits backward should be smaller than the span for digits forward, according to this line of reasoning, because the operation of reversing the digits takes a certain amount of time, during which the information in working memory decays. And indeed, backward digit recall is consistently inferior to forward digit recall. Subjects who can recall seven digits forward can usually recall only five digits backward. Beyond some optimal point, which varies across individuals (the average being seven digits), the greater the number of digits presented, the smaller the number of digits recalled in correct order. Both forward and backward digit span are correlated with psychometric *g*, and are often included in IQ tests such as the Stanford–Binet and Wechsler scales. Yet backward digit span, because of its greater processing demands, consistently shows a higher *g* loading than forward digit span.

Similarly, successful performance on all mental test items depends on various elementary cognitive processes, the more complex items making the greater processing demands in terms of information storage, operations performed, information retrieved from LTM, and so forth. The more complex the information and the operations required, the more processing time demanded, and consequently, the greater the advantage of speed in all the elementary processes involved. Loss of information due to overload interference and decay of traces that were inadequately encoded or rehearsed for storage or retrieval results in "breakdown" in grasping all the essential relationships required for arriving at the correct answer. Speed of information processing, therefore, should be increasingly related to success in dealing with cognitive tasks as the informational load increasingly strains the individual's limited working memory. Thus, the most discriminating test items are those that "threaten" the processing system at the threshold of "breakdown," beyond which erroneous responses occur. In a series of test items of graded complexity, this "breakdown" would occur at different points for various individuals. If individual differences in the speed of the elementary components of information processing can be measured in tasks that are so simple as to rule out "breakdown" failure, moreover, it should be possible to predict the individual differences in the point of "breakdown" for more complex tasks, such as Raven Matrices items or other items typically found in IQ tests. This is the hypothesized basis for the observed correlations between response latencies on ECTs and scores on complex *g*-loaded tests.

Such correlations will differ in magnitude because of the complexity of the ECTs and of the test items themselves, since the more complex items involve a greater number of different processes, allowing more shared variance. As is well known in factor analysis, complex tasks are more highly *g*-loaded than simple tasks. Hence, a variety of ECTs combined will show a larger correlation with psychometric *g* than will any single ECT, however reliably the response latencies are measured. Correlations between ECTs and psychometric tests may also be limited by the degree to which successful performance on the tests depends upon specific knowledge content or learned strategies for solving certain types of problems (e.g., the use of Venn diagrams for solving syllogisms); these correlations may also be limited by the extent to which individuals differ in possessing such knowledge or skills. Some of the variance in psychometric test scores – just how much is still uncertain – is attributable to various "metaprocesses." Such metaprocesses include strategies for selecting, combining, and using elementary processes, problem recognition, rule application, planning, allocation of resources, organization of information, and monitoring one's own performance. Different metaprocesses are intercorrelated because they have certain elementary processes in common, because they all must operate within the time constraints of working memory, and also because the experiential factors that inculcate certain metaprocesses are correlated in the educational and cultural environment.

## Testing Spearman's hypothesis

Spearman's hypothesis that the magnitudes of black–white mean differences on various mental tests are directly related to the tests' *g* loadings, if fully substantiated, would be an important and unifying discovery in the study of population differences in mental abilities. Spearman's hypothesis, if true, would mean that the black–white difference in test scores is not attributable merely to idiosyncratic cultural or linguistic peculiarities in this or that test, but to a general factor which all mental tests measure, and which some tests measure to a greater degree than others.

The finding of mean differences in *g* between populations, of course, does not necessarily rule out cultural influences (e.g., those lowering its reliability, or its validity relative to external criteria). But *g* would reflect only those broad influences which are manifested not in any particular item, test, or type of test but in a very wide variety of tests that differ greatly in the types of knowledge and cognitive skills that they sample.

No data, so far, have been collected specifically for the purpose of testing Spearman's hypothesis. However, a search of the psychometric literature for relevant data has turned up 11 large-sample studies containing appropriate data that may be analyzed to determine whether the results are predominantly consistent or inconsistent with Spearman's hypothesis.

For the sake of precision, Spearman's hypothesis should be stated in two forms that can be termed *strong* and *weak*, respectively, although Spearman himself did

not suggest this distinction. The strong form of the hypothesis holds that the magnitudes of the black–white differences (in standard score units) on a variety of tests are directly related to the tests' *g* loadings, because black and white populations differ only on *g* and on no other cognitive factors. The weak form of the hypothesis holds that the black–white difference in various mental tests is predominantly a difference in *g*, although the populations also differ, but to a much lesser degree, in certain other ability factors besides *g*.

## Methodological desiderata

The most obvious test of Spearman's hypothesis would be to calculate the correlation between the *g* factor loadings of various tests and the mean black–white differences (in standardized units) on the various tests.

*Factor analysis* and *principal components* are distinct, but rather closely related, mathematical models for transforming a matrix of intercorrelated observed variables into a set of underlying variables, of which the observed variables are linear functions. In principal component analysis, the derived variables (termed principal components) are *orthogonal* (i.e., uncorrelated). In factor analysis, the derived latent variables (termed factors) may be *either* orthogonal or oblique (i.e., correlated with one another). In principal components, the *n* observed variables are transformed into *n* linearly independent variables, or components, which account for the total variance in the observed variables, with the first principal component accounting for the largest proportion of the total variance, the second principal component accounting for the second largest proportion, and so on to the *n*th component, which accounts for the smallest proportion of the variance. In factor analysis, the total variance of the observed variables is divided into two main portions: (1) a number of *common factors* (which empirically are always fewer than the number of observed variables) and (2) a residual variance, consisting of *specificity* (i.e., that portion of the reliable or true-score variance of each observed variable which is not shared by any of the other observed variables in the analysis) and *error variance* due to errors of measurement, or unreliability. The common factors are latent variables shared by two or more of the observed variables. An observed variable's *communality* is that proportion of its variance which is attributable to common factors.

The largest common factor (i.e., the factor accounting for the largest proportion of the total variance attributable to all of the common factors) may often be interpreted as a general factor, or *g*. (Also, the first principal component is often loosely termed a "general factor.") The mathematical basis of principal components and common factor analysis is succinctly explicated by Kendall and Stuart (1976, Chap. 43). More detailed treatments can be found in books by Cattell (1978), Harman (1967), and Mulaik (1972).

There are three main methods currently in use for factoring a correlation matrix. Each method yields the general factor of a collection of tests: the first principal component, the first principal factor, and a second-order *g* factor derived from a hierarchical factor analysis, that is, the general factor among the obliquely rotated first-order factors. For the data under consideration here, it turns out that all three methods yield such similar results that findings and conclusions are essentially the same. In fact, *g* loadings have been extracted by all three methods in the present study. The Burt–Tucker (Cattell 1978, pp. 251–55) coefficient of congruence[2] applied to the *g* factor loadings extracted by each of the three methods shows values ranging from .990 to .999. This is a typical finding (e.g., Silverstein 1980a; 1980b). However, because the first principal factor is the most generally preferred representation of *g* among experts in factor analysis, results reported in the present paper are generally based on the first principal factor. In two studies for which other factors besides *g* are also of theoretical interest, however, the hierarchical second-order *g*, obtained by the Schmid–Leiman (1957) orthogonalization transformation, is used.[3] (The Schmid–Leiman hierarchical factor analysis differs from the more familiar Thurstone hierarchical factor analysis in that the Schmid–Leiman analysis residualizes the oblique [correlated] primary factors, and thereby orthogonalizes them. This procedure makes the primary factors smaller, since their common variance, which now exists in the factors at the next higher level of the hierarchy, has been removed. Orthogonalization is similarly applied at each higher level of the hierarchy, so that all the factors within levels and between levels of the hierarchy are made orthogonal to one another, and each of the original variables [tests] is projected onto each of the orthogonal factors at each level of the hierarchy. A distinctly different alternative method of hierarchical factor solution that achieves a result which is identical to that of the Schmid–Leiman procedure has been developed by Wherry, 1959.) For the present data, the congruence coefficients between the Schmid–Leiman *g* and the first principal factor are greater than +0.99 in both the black and white samples.

It should be understood, of course, that the first principal factor of any given collection of mental tests (or other measurements) does not necessarily represent the same general factor as Spearman's *g*, or the same general factor that would be extracted from some quite different collection of tests. It turns out, however, that different batteries of tests, provided they comprise a considerable number and diversity of tests, do, in fact, yield highly similar *g* factors (Jensen 1980a, pp. 233–34). That is to say, the sets of *g* factor scores derived from the different test batteries administered to the same subject sample are highly correlated with one another. Moreover, examination of the nature of the tests showing the highest *g* loadings in any battery usually reveals that the items in these most *g*-loaded tests formally reflect Spearman's characterization of *g* as the capability for abstract reasoning, or, to use Spearman's own words, "the eduction of relations and correlates." The inferential ability reflected in highly *g*-loaded test items has presumably operated either largely in the person's past (as in the acquisition of vocabulary and general information [Sternberg & Powell 1983; Werner & Kaplan 1952]), or largely in the immediate test situation itself (as in solving novel figure analogies or progressive matrices). Cattell (1963) has characterized these two aspects of *g* as *crystalized* and *fluid* intelligence, or $g_c$ and $g_f$, respectively. In native-born, English-speaking subpopulations in the United States, there is generally a very high correlation between $g_c$ and $g_f$– so

high, in fact, that these two facets of general intelligence cannot always be clearly distinguished by factor analysis.

The average difference between two groups on a given test must, of course, be expressed in standardized units, if it is to be meaningfully compared with the average group difference on some other test. I have used as the standardized unit the square root of the variance *within* groups, referred to henceforth as a sigma ($\sigma$) unit. This $\sigma$ unit is equivalent to the weighted average standard deviation within groups, the weights being the respective sizes of the two samples. That is, the sigma unit for two groups, A and B, would be

$$\sigma = [(N_A\sigma^2{}_A + N_B\sigma^2{}_B)/(N_A + N_B)]^{1/2}$$

where $\sigma_A$ and $\sigma_B$ are the standard deviations of groups A and B, respectively, and $N_A$ and $N_B$ are the numbers of persons in each group. The mean difference between the groups expressed in $\sigma$ units is simply $\bar{d}_\sigma = (\bar{X}_A - \bar{X}_B)/\sigma$.

One kind of evidence supporting the Spearman hypothesis, then, would consist of a positive coefficient of correlation (or other index of relationship) between the *g* loadings of specific tests and the standardized mean black–white difference ($\bar{d}_\sigma$) on these tests. Since the correlation would usually be based on a small *N* (i.e., the number of tests), the magnitude of such correlations and their consistency across different samples of persons and different batteries of tests should take precedence over the level of statistical significance of any single correlation as evidence for Spearman's hypothesis. Because the *g* loadings derived from a particular battery of tests are not statistically independent of one another and do not qualify as a random sample from a population with an assumed normal distribution, and because the same is true of the standardized mean black–white differences on the tests, the Pearson product–moment coefficient of correlation (*r*) between *g* loadings and mean differences, although it is the most precise index of the degree of linear relationship between the two sets of variables, cannot, in a strict sense, be tested for statistical significance. Therefore, significance tests are not here applied to the Pearson *r* when used as an index of relationship between *g* loadings and mean black–white differences. However, in addition to the Pearson *r*, the corresponding Spearman rank-order correlation, $\rho$, is also reported, because its level of significance does not rest on any assumptions about the distributional characteristics of the two variates (Kendall & Stuart 1976, pp. 494–99). As a nonparametric, or distribution-free, test of independence or index of relationship, the rank correlation's level of significance is simply the proportion of all possible *n*! permutations of the *n*-ranked pairs of variables for which the absolute value of $\rho$ is equal to or greater than the obtained $\rho$.

Ideally, four methodological caveats should be observed in investigating Spearman's hypothesis.

1. Factor analysis should be performed in the two population groups separately, so that the factor loadings (via the zero-order correlations from which the factors are derived) are not contaminated by population differences on the various tests. If the same factors are found in both populations, it is appropriate to use the factor analysis of whichever sample is larger, because this analysis will have the higher reliability. The first principal factor, or *g*, in a battery of tests must be essentially the same factor, within the limits of sampling error, in both populations. This requirement can be tested as follows: We determine the degree of similarity between populations in the *pattern* of factor loadings over the various tests by obtaining the congruence coefficient, $r_c$, between the two sets of loadings. A high congruence coefficient (i.e., at least .90) means that the magnitudes of the factor loadings on the various tests are highly similar for both populations. A potential problem arises if all the tests are nearly equally loaded on *g*. In this event, because of random sampling error, the slight differences in *g* loadings may not form a sufficiently reliable pattern to allow a substantial correlation between the population groups. The split-half reliability of the pattern of *g* loadings can be estimated by splitting the subject sample into random halves and factor-analyzing each half. The correlation between the factor loadings of the two halves, boosted by the Spearman–Brown prophecy formula [boosted $r = 2r_{hh}/(1 + r_{hh})$, where $r_{hh}$ is the correlation between the half-sample profiles], gives an estimate of the reliability of the pattern of *g* loadings for the total sample. The reliability of the pattern of mean group differences on the various tests can be estimated by the same procedure. The correlation between the pattern of *g* loadings and the pattern of group differences can then be corrected for attenuation in the usual way, by dividing the correlation by the geometric mean of the two reliability coefficients.

2. The population samples being compared should not have been selected in terms of any highly *g*-loaded criterion. For example, we could not properly test Spearman's hypothesis by using black and white students in a highly 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. Hence, any selection of subjects on general ability would work directly against the Spearman hypothesis to some degree. What is more, the *g* factor extracted from tests given to highly 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.

3. Test reliability affects both factor loadings and group mean differences (in $\sigma$ units). Both variables are attenuated by measurement error. If, therefore, reliability differs markedly from one test to another, the correlation between the profile of the tests' *g* loadings and the profile of the mean population differences on the tests will be spuriously inflated by the common influence of unreliability (measurement error) on both variables. This is probably not a serious drawback if all the tests have quite high and similar reliabilities or if there is no systematic relationship between tests' reliabilities and their intrinsic *g* loadings (i.e., the *g* loadings after correction for attenuation). The importance of these possibilities must be empirically investigated. Of course, it is always most desirable, when the test reliabilities are known, to correct both the *g* loadings and the mean differences for attenuation. This is accomplished by dividing each variable (i.e., the *g* loading and the mean difference) by the square root of the test's reliability coefficient. The studies reviewed here have provided only internal consistency reliabilities (KR-20 or split-half), and these have been used to correct the *g* loadings and mean differences for attenuation. It

would also have been desirable to correct for attenuation based on test–retest reliability (i.e., temporal stability of test scores), but these reliabilities were not available. Although the two types of reliability are conceptually distinct, empirically they are usually quite similar for mental tests.

4. Some caution must be exercised in the theoretical interpretation of a high correlation between tests' *g* loadings on the first principal factor and the mean differences between groups. Not every collection of tests necessarily yields a first principal factor that can be properly interpreted as Spearman's *g* in the psychological sense intended by Spearman. The first principal factor is affected by variation in psychometric sampling; different collections of tests will result in somewhat different first principal factors, especially if each collection has a concentration of highly similar tests that differ quite markedly from the tests in other collections. Thus one must look for evidence that the first principal factor can reasonably be interpreted as Spearman's *g*. Marker tests with known high *g* saturations, as evidenced by other factor analytic studies, may serve as an important indicator. Another potent indicator is the degree of relationship between the profile of various tests' *g* loadings and the profile of these same tests' correlations with IQ or total scores on the best tests of general intelligence in terms of their validity for predicting performance in educational, occupational, and other practical criteria. It seems safe to say that most of the variance (probably as much as 75% to 85%) in total scores on standard omnibus intelligence tests represents Spearman's *g*. Therefore, the loadings on the first principal factor of a collection of cognitive tests should be quite highly related to the correlations of these tests with total scores on tests of general intelligence or IQ if the first principal factor is to be properly interpreted as Spearman's *g*. Finally, of course, we should inquire as to the nature of the two or three tests that show the highest loadings on the first principal factor of our collection of tests, in order to see if these highly loaded tests display the properties of inference or relation eduction, abstractness, and transformational complexity that best characterize Spearman's *g* psychologically.

If the psychological interpretation of the first principal factor (as contrasted with its purely mathematical interpretation) is in doubt, then what would be the meaning of a high degree of relationship between the factor loadings (derived within either black or white samples) of the various tests and the sizes of the black–white mean differences on those tests? If there is a doubt that the first principal factor is very similar to Spearman's *g*, such a relationship could, of course, neither confirm nor disconfirm Spearman's hypothesis. However, such a finding would mean, at the very least, that whatever linear composite of these various tests discriminates the most among individuals *within* each population also discriminates the most *between* the means of the two populations. This condition implies, of course, that individual differences *within* the populations and the mean difference *between* the populations are factorially the same or highly similar, whatever the psychological nature of the factor may be. In other words, the first principal factor of this battery of tests discriminates between black and white individuals on the same basis as it discriminates between individuals in the same population, whether or not the first principal factor is psychologically interpretable as Spearman's *g*. This would be the expected outcome, of course, if the tests in the battery were not biased in discriminating individual differences.

Elsewhere (Jensen 1980a), I have pointed out an alternative interpretation of the empirical findings which Spearman's hypothesis attempts to comprehend:

> Blacks and whites differ merely in overall level of performance on all test items (i.e., there is no race × items interaction), and those items (or subtests) that contribute the most to the true-score variance (by virtue of high reliability and optimal difficulty level) among individuals of either race thereby also show the largest mean differences between the races, and they are also the most heavily loaded on a general factor (i.e., the first principal component) that, by its mathematical nature, necessarily accounts for more of the variance than any other factor, regardless of the psychological nature of the first principal component extracted from the particular collection of tests. By this interpretation, the only condition needed to yield results at least superficially consistent with Spearman's hypothesis is that there be no appreciable race × items or race × tests interactions or, in other words, that the tests not be racially biased. (Pp. 548–49)

Not only does this explanation now appear far too superficial, it is seriously inadequate on at least two counts. In the first place, as is shown by the evidence in the present article, there is a correlation between black–white differences and *g* loadings on various tests, even when differences in test reliability are taken into account by correcting the *g* loadings and the mean differences for attenuation (i.e., unreliability). Second, tests and single items still show differences in *g* loadings when they are equated in difficulty level and variance; that is, tests' or items' *g* factor loadings and differences in factor loadings are not mere artifacts of differences in variance or level of difficulty, and *g*, or the first principal factor, is not explainable in terms of these variables. Certain types of items and tests, whose common characteristics cannot be described in terms of information content or surface appearance alone, have larger *g* loadings more consistently than other items or tests. Even the same test can take on different *g* loadings under different degrees of what might be termed "cognitive strain." We see this most clearly in dual tasks (or competition tasks) in which the subject is required to perform two different elementary cognitive tasks, either simultaneously or in immediate succession. Dual tasks can be used for measuring storage/processing trade-off in working memory. The more of the capacity of working memory that is used for short-term storage of information, the less capacity there is available for other forms of information processing – encoding, discrimination, transformation, and so on. Consequently, a dual task puts a greater strain on the storage and processing capacity of working memory. In a dichotic listening task, for example, a person simultaneously hears a different pattern of three tones in each ear (e.g., left ear: high, low, high; right ear: low, high, low) and is then randomly postcued to report the pattern presented to one ear. Stankov (1983) has made the discovery that performances on a variety of ECTs are more highly intercorrelated, and are therefore more heavily *g*-loaded, when they are presented in the dual-task para-

digm than when presented as single tasks. Also, Stankov distinguishes between the *active* and *passive* aspects of working memory, terms corresponding to the *processing* and *storage* of information, and concludes that the active component of working memory is more highly correlated with fluid g than is the passive component: "operations performed on information in working memory are more indicative of fluid intelligence than is the ability to hold this information in working memory" (Stankov 1983, p. 51). This observation is very similar to the distinction between Level I and Level II abilities as *encoding and retention* of stimulus input (Level I) and *mental manipulation* of encoded material (Level II) (Jensen 1974a). The distinction between Level I and Level II abilities was originally suggested by the finding that blacks and whites differed, on average, very much less on Level I than on Level II types of tests. (This evidence has been most extensively reviewed by Vernon, 1981a.) Also, in a factor analysis of reaction times obtained on eight ECTs, Vernon and Jensen (1984) found that dual tasks consistently showed larger loadings on the first principal factor than did component tasks when administered singly. In this same study, moreover, the largest average black–white differences (in σ units) in RT occurred on the dual tasks, a finding clearly consistent with Spearman's hypothesis. It is for such reasons that Spearman's hypothesis cannot be dismissed as reflecting only psychometric or factor analytic artifacts. We now know at least one of the conditions by means of which tests' g loadings can be manipulated experimentally. Such manipulation does not necessitate altering the information contents of tests or their specific skill requirements. These g loadings can be increased or decreased simply by varying the demand load placed on the information-processing capacities of individuals being tested.

## Evidence for the Spearman hypothesis

The recent literature (since 1970), including doctoral dissertations and government reports, was searched for data that meet the basic requirements for testing the Spearman hypothesis: batteries of six or more diverse tests administered to large black and white samples that were not highly selected on intelligence, and, in addition, presentation of the intercorrelations among all the tests as well as their means and standard deviations in the black and white samples. All of the nearly 500 references in the exhaustive compendium of studies in this area by Osborne and McGurk (1982) were considered. Eleven studies were found suitable for analysis. (These studies are listed and summarized in the Appendix.) Within each test battery, g factor was extracted separately for black and white scores, and the factorial similarity was measured by the congruence coefficient. (In one study, correlations were not available for both black and white samples separately but only for a predominantly white sample.) Also, the mean difference between the black and white samples on each of the tests was calculated in standard score units. In 7 of the 11 studies, the reliabilities of the tests were available, permitting corrections for attenuation of the g loadings and of the mean black–white differences on each test. For each study, a Pearson correlation was then obtained between the g loadings and the mean black–white differences. (As protection against possible outliers among g loadings or differences that might tend spuriously to inflate or deflate the Pearson correlation, Spearman's rank correlation [ρ], corrected for tied ranks, was also computed: the ρ seldom deviates appreciably from the Pearson *r*, however. Also, as explained previously, because the sampling distribution of the Pearson *r* between g loadings and mean group differences is not known, a test of significance of the correlation between g loadings and black–white differences can be strictly applied only to Spearman's rank correlation, rho (ρ).

Figure 1 shows the correlation scatter diagram relating the average black–white differences to the g loadings on all 121 of the tests in the 11 studies. (No corrections for attenuation have been applied here.) This bivariate distribution clearly reveals that there is considerable variation both in g loadings and in the size of the black–white difference as expressed in standard score units (*z* or σ units). It is true here, as it is in each of the separate studies, that the g loadings show considerably less variability (as measured by the coefficient of variation) than do the mean differences, $\bar{D}$. The main reason for this is probably that the tests in most of these batteries were selected by their authors because they are rather good measures of general ability, and so there are few tests with very low g loadings. This restriction of range in the g loadings, of course, tends to lower the correlation between the g loadings and the mean differences. By the same token, according to Spearman's hypothesis, the variability of the mean differences on the various tests should also be more constrained than would be the case if

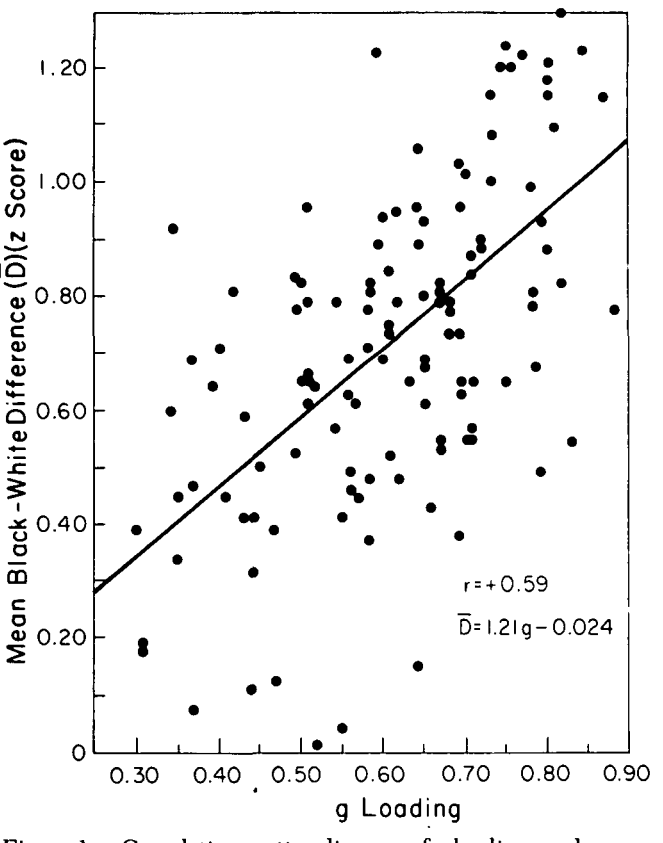


Figure 1. Correlation scatter diagram of g loadings and mean black–white differences (in standard score units) for 121 tests in 11 studies. The g loadings and differences are not corrected for attenuation.

the tests' g loadings were more heterogeneous. To get some idea of how dependent the $g \times \bar{D}$ correlation is on the variability of g loadings and the variability of $\bar{D}$ in each battery of tests, it turns out that the standard deviations of the g loadings and of $\bar{D}$ in each study predict the correlation between g and $\bar{D}$ over the 11 studies with a multiple $R$ of 0.73. This means that about half of the variation among the correlations obtained between g loadings and differences within each of the 11 test batteries is attributable to psychometric artifacts of the particular batteries (viz., too little heterogeneity of the tests with respect to g). These artifacts are irrelevant to Spearman's hypothesis and, in varying degrees, attenuate or obscure its manifestation in different collections of tests. Hence, with some optimal degree of heterogeneity of g loadings in a large battery of tests, it seems a reasonable conjecture that the correlation between g loadings and mean black–white differences would be much higher than the overall correlation of +0.59 found in the present data.

Despite the attenuation of correlation between g and $\bar{D}$, we see in Figure 1 that Spearman's hypothesis, at least in its weak form, is substantiated at a high level of statistical significance by these 11 studies. The Pearson $r$ between g loadings and average black–white differences is +0.59 (Spearman's $\rho = +0.59$, $p < .001$).

But why is the correlation not higher than this, if Spearman's hypothesis is true? Besides the restriction of range, which I have already noted as the main cause of attenuation of the correlation, there is the fact that the g factor is not exactly the same g in every collection of tests. Remarkably, there is a generally high positive correlation between the g obtained in any one battery of tests and the g obtained in any other battery. Yet the g factor is not a constant across all batteries of tests; it is determined in part by the nature of the particular combination of tests making up the whole battery from which it is extracted. In other words, not every g is an equally good g.

As yet we have no single objective criterion of what constitutes a good g. It would seem reasonable to assume, however, that all of these 11 test batteries yield estimates of g that approximate a good g to varying degrees. Until we have discovered the essential nature of g in terms that are independent of factor analysis, we cannot objectively claim that the g of any one battery is necessarily a better g than the g of any other battery. The solution to this problem is one of the major challenges facing cognitive psychologists. The fact that the gs of all the batteries of diverse cognitive tests are similar does suggest the possibility that there is a theoretically true g toward which the obtained gs from various test batteries tend to converge, although this point remains controversial at our present state of knowledge.

The robustness of the Spearman hypothesis is shown by the finding that in every one of the 11 test batteries there is a positive correlation (with a unit-weighted average of +0.60) between g loadings and the mean black–white difference.

Also consistent with the Spearman hypothesis is the finding that the regression line (see Figure 1), if extended to the point at which the g loading is zero, indicates a mean black–white difference that is also very close to zero ($-0.024\sigma$, to be exact). If the regression line is extended up to the point at which the g loading is 1.00, the mean black–white difference is approximately $1.21\sigma$, which is about the upper limit of the difference actually found for any test on representative samples of the black and white populations. Hence, the total range of actual black–white mean differences does not extend beyond the range that would be theoretically predicted by the lowest and highest positive g loadings that any test could possibly have (i.e., 0 and 1).

Averaging over all 11 studies, we can compare the overall mean black–white difference on the tests having the highest and the lowest g loadings in each battery, with the results shown in Table 1. The differences, as indicated by the correlated $t$ test, are significant well beyond the .001 level, even with only 9 degrees of freedom. Although the precise meaning of this significance level may be questioned because it is based on contrasting the black–white difference on the single most and the single least g-loaded tests in each battery, it should be noted that a *nonsignificant* $t$, in this case, would definitely warrant rejection of Spearman's hypothesis.

Table 2 shows the highest and lowest g-loaded tests in each study, with the corresponding g loadings and the mean black–white differences, $\bar{D}$.

There is also evidence that Spearman's hypothesis holds not only for the various tests factor-analyzed within a given battery but also for the overall average of the black–white differences on the tests in each battery in relation to the average of the g loadings of the tests in each battery. The correlation between the average black–white differences and the average g loadings across the 11 studies is +0.54 ($\rho = +0.42$, $p < .05$). This is shown in Figure 2. Some of the variation in the mean black–white difference in various studies is associated with the variation in g loadings (and the correlated variation in the black–white differences) among the tests in each battery. These theoretically irrelevant sources of variance merely attenuate the manifestation of Spearman's hypothesis. If we partial out the effects of variation in g loadings and variation in mean differences (i.e., the standard deviation of the g loadings and the standard deviation of the differences, in each battery), the resulting second-order

Table 1. *Mean difference (in σ units) between black and white samples on tests with the highest and lowest g loadings in each of 11 studies*

| | Highest g | | Lowest g | | Difference | | Correlated |
|---|---|---|---|---|---|---|---|
| | Mean | *SD* | Mean | *SD* | Mean | *SD* | *t* Test |
| Black g | .997 | .528 | .593 | .517 | .404 | .181 | 7.08* |
| White g | .948 | .531 | .554 | .491 | .395 | .165 | 7.57* |

*$p < .001$

Table 2. *Tests with highest and lowest g loadings[a] (corrected for attenuation) in each of 11 batteries and the mean black–white difference, $\bar{D}$ (in σ units)*

| Number of tests | Highest g | g | $\bar{D}$ | Lowest g | g | $\bar{D}$ | Study |
|---|---|---|---|---|---|---|---|
| 13 | WISC-R Vocabulary | .78 | .88 | WISC-R Tapping Span | .39 | .33 | Jensen & Reynolds (1982) |
| 12[b] | WISC-R Vocabulary | .78 | .67 | WISC-R Coding | .30 | .39 | Reynolds & Gutkin (1981) |
| 12 | WISC-R Information | .79 | .93 | WISC-R Coding | .39 | .50 | Sandoval (1982) |
| 12 | WISC-R Vocabulary | .76 | .90 | WISC-R Coding | .41 | .46 | Mercer (1984) |
| 12 | SAT–Verbal | .87 | 1.15 | Mosaic Test | .34 | .92 | Nat. Long. Study |
| 13 | WRAT–Arithmetic | .71 | .55 | WISC-R Coding | .37 | .17 | Nichols (1972) |
| 10 | ASVAB–Arithmetic | .91 | 1.16 | ASVAB–Coding Speed | .56 | .96 | Dept. of Defense (1982) |
| 8 | GATB–Form Perception | .82 | .55 | GATB–Manual Dexterity | .43 | .08 | Dept. of Labor (1970) |
| 13 | K-ABC Arithmetic | .88 | .82 | K-ABC Gestalt Closure | .56 | .39 | Kaufman & Kaufman (1983) |
| 6 | Sentence Completion | .81 | .82 | WISC-R Backward Digit Span | .68 | .41 | Veroff et al. (1971) |
| 10 | Reading Comprehension | .75 | .65 | Spatial Reasoning | .31 | .19 | Hennessy & Merrifield (1976) |

[a]The g loading derived from the larger sample (black or white) was used in this analysis. [b]Black and white samples in this study (Reynolds & Gutkin 1981) are matched on socioeconomic status.

partial correlation between the mean g loadings and mean black–white differences is +0.85, which impressively bears out Spearman's hypothesis. No other characteristic of the tests (e.g., other factors besides g, or characteristics such as verbal, nonverbal, performance, oral or paper-and-pencil, individual or group administered) is as systematically related to the size of the black–white differences as the tests' g loadings. The tests with the lowest g loadings, for the most part, appear to measure short-term memory, clerical speed and accuracy, and manual dexterity, all abilities that involve very little relation eduction, which Spearman regarded as the strongest manifestation of g.

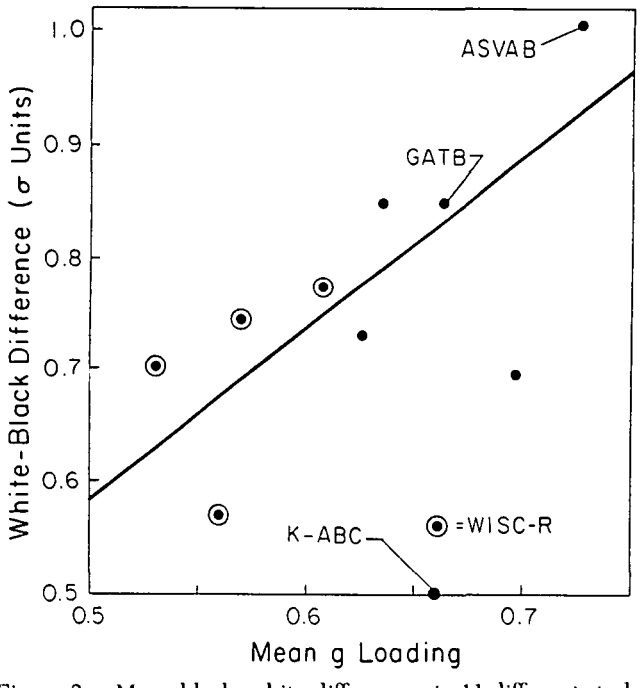


Figure 2. Mean black–white differences in 11 different studies as a function of mean g loading of tests in each study. (The regression of the mean difference [$\bar{D}$] on mean g loading is $\bar{D} = 1.38g - .11$.) In the one study (Reynolds & Gutkin 1981) of the WISC-R that falls below the regression line, the black and white samples were matched on four demographic variables, including socioeconomic status.

## Strong and weak forms of Spearman's hypothesis

A study of the national standardization sample of the WISC-R (Jensen & Reynolds 1982), based on 1,868 white and 305 black children, bears out Spearman's hypothesis but contradicts it in its strong form, because significant, but small, black–white differences were found on other factors besides g. When the WISC-R is subjected to a Schmid–Leiman hierarchical factor analysis, it yields four factors that are virtually identical for both populations: g, verbal, spatial, and memory. When factor scores on each of these four factors are computed for every black and white subject, the populations show significant mean differences on all four factors, a finding that contradicts the strong form of Spearman's hypothesis. But the weak form is strongly upheld, as the g factor accounts for more than seven times as much of the between-population variance as the other three factors combined. Black testees exceed white testees on the Memory factor (0.32σ), whereas white testees exceed black testees on the g (1.14σ), Verbal (0.20σ), and Performance (0.20σ) factors.

The same data exhibit another contradiction of the strong form of Spearman's hypothesis, as shown in Figure 3. The point-biserial correlation of each WISC-R subtest with population (coded as black = 0, white = 1) represents the degree to which the populations differ on each test, with zero correlation representing zero difference, and positive correlations indicating white superiority; negative correlations, black superiority. The upper profile in Figure 3 shows the raw correlations. The lower profile shows the correlations with Full Scale IQ partialed out. This, in effect, equates the black and white samples on g (since IQ is correlated .98 with the g factor scores of the WISC-R), permitting us to see whether and how the black and white groups differ on the subtests after their difference on g is removed. It can be seen that the populations still differ significantly on 6 of the 13 subtests, with black performance superior on Arithmetic and Digit Span (which are loaded on the Memory factor) and white performance superior on Comprehension, Block Design, Object Assembly, and Mazes (the last 3 subtests measure predominantly spatial ability, in addition to g).

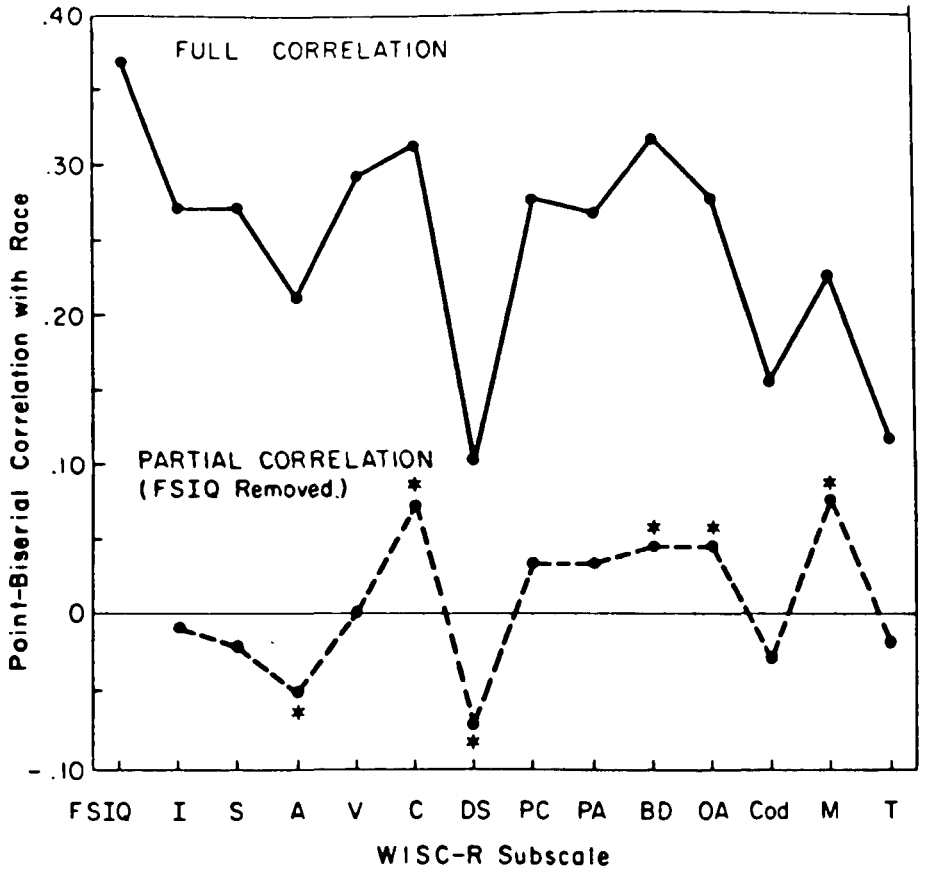


Figure 3. Point-biserial correlation as an index of black–white difference on FSIQ and on each of 13 subtests of the WISC-R (Wechsler Intelligence Scale for Children–Revised). The upper profile shows the actual group differences. (All are statistically significant.) The lower profile shows the black–white differences on the 13 subtests after FSIQ has been partialed out, in effect equating the population groups on general intelligence. Those differences which are significant beyond the 0.05 level are indicated by asterisks. I —Information, S—Similarities, A—Arithmetic, V—Vocabulary, C—Comprehension, DS—Digit Span, PC—Picture Completion, PA—Picture Arrangement, BD—Block Designs, OA—Object Assembly, Cod—Coding [Digit Symbol], M—Mazes, T—Tapping [Knox Cubes]. (From Jensen & Reynolds 1982.)

It is noteworthy that, with *g* held constant, there is no black–white difference on Vocabulary. Another important point was reported in the same study: When profiles are created by the same method to show the IQ-partialed correlations between WISC-R subtests and the children's socioeconomic status (separately within each population), the profiles are extremely different from the black–white profile; in fact, the two social status profiles are *negatively* correlated (−0.63 for black children and −0.45 for white) with the black–white population profile (i.e., the lower profile in Figure 3). This means that with IQ held constant, the pattern of black–white subtest differences is quite different from the pattern of subtest differences associated with high and low socioeconomic status.[4] This finding flatly contradicts the notion that the pattern of black–white differences in test performance merely reflects the overall black–white difference in socioeconomic status.

Another way of looking at the Spearman hypothesis is shown in Figure 4 for the WISC-R standardization data. (Details of this study are reported in Jensen & Reynolds 1982.) The rank-order correlation between the *g* and $\bar{D}$ profiles is +0.75 ($p < .01$); the Pearson $r$ = +0.73. Because data on individual subjects were available in this study, it was possible to obtain the split-half reliabilities of the profiles of *g* loadings and differences and to correct the correlation between the profiles for attenuation due to sampling error. The disattenuated Pearson $r$ is +0.81. Black–white differences were also expressed as point-biserial correlations; their correlation with the *g* loadings is negligibly different from the correlation of *g* with the mean black–white differences in $\sigma$ units, which is hardly surprising, as the point-biserial $r$ is virtually a linear function of the standardized mean differences in the range of differences less than $2\sigma$. (For further details of this analysis, see Jensen & Reynolds 1982, pp. 433–35.)

The fact that the WISC-R measures other group factors besides *g*, on which black and white populations also differ in varying degrees, tends to attenuate the correlation between *g* loadings and the mean black–white differences on the 13 subtests. Only if the strong form of Spearman's hypothesis were true would this *not* be the case. If we eliminate the differential effects of the verbal and performance factors by testing Spearman's hypothesis just within the set of the 6 verbal subtests, the correlation between *g* loadings and mean black–white differences rises to +0.86; for just the 7 performance subtests the correlation is also +0.86, again substantiating the weak form of Spearman's hypothesis.

It should not be assumed, however, that any two groups that differ because of cultural or linguistic deprivation of one group relative to the other will, of mathematical necessity, show a correlation between the *g*-loadedness of various tests and the magnitudes of the

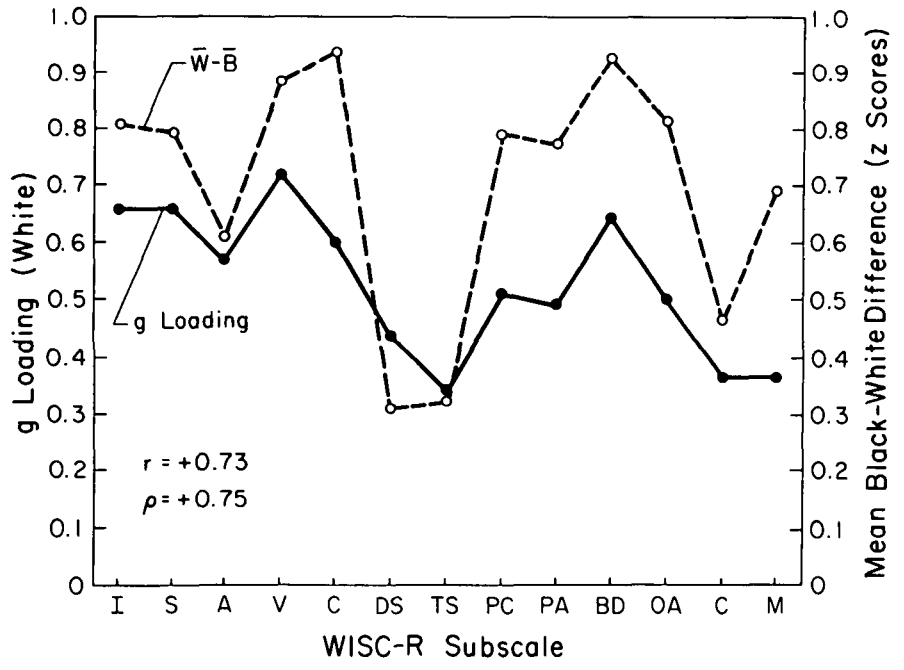


Figure 4. Mean black–white differences and g loadings on 13 WISC-R subtests. The correlation between the profiles of the black–white differences and g loadings is indicated by the Pearson r and the Spearman ρ (rank-order correlation).

group differences. Quite different results emerge in a comparison of congenitally or preverbally deaf children and normal-hearing children on the WISC-R (Braden 1981). Because the verbal tests of the WISC-R are inappropriate for the congenitally and preverbally deaf, only the nonverbal Performance scales were used. (The deaf sample of 1228 children is described by Anderson and Sisco, 1977, and Sisco, 1982.) The profile of average differences between hearing and deaf children can be compared with the profile of average differences between black and white children in the WISC-R national standardization sample, and with the profile of g loadings based on factor analysis of just the six Performance subtests in the standardization sample ($N$ = 2200). The results are shown in Figure 5. The Pearson correlation between g loadings and the mean black–white differences is +0.97;

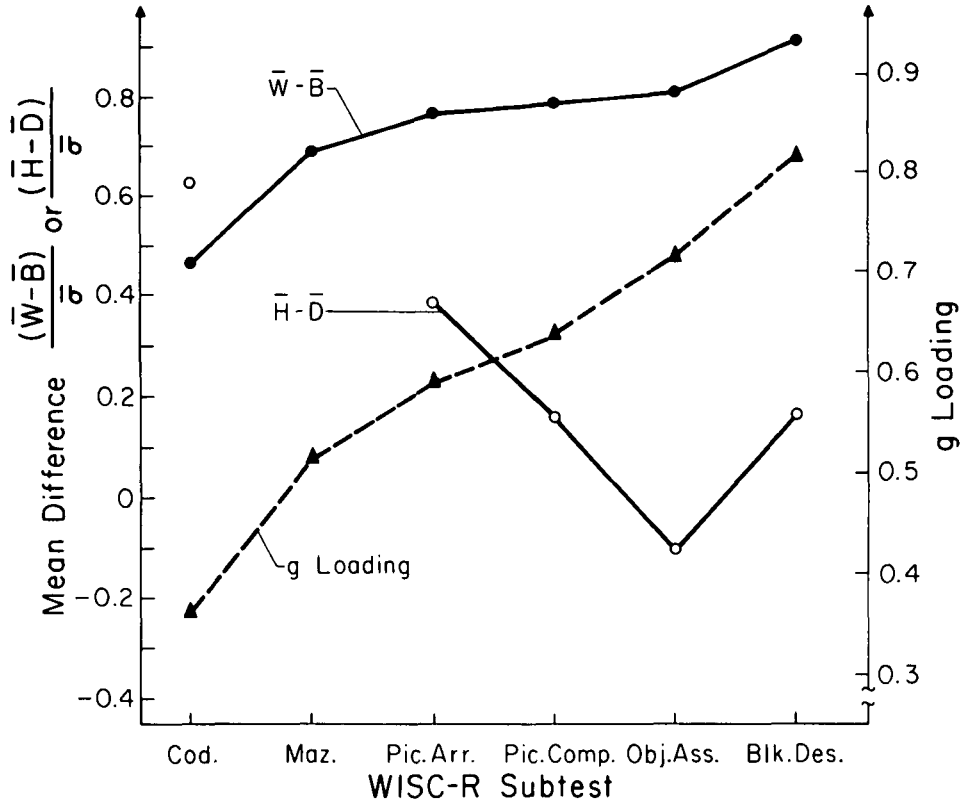


Figure 5. Average differences between hearing and deaf children and black and white children, and g loadings of WISC-R Performance subtests. (Note: The Mazes subtest was not obtained in the deaf sample.) (Based on data from Braden 1984.)

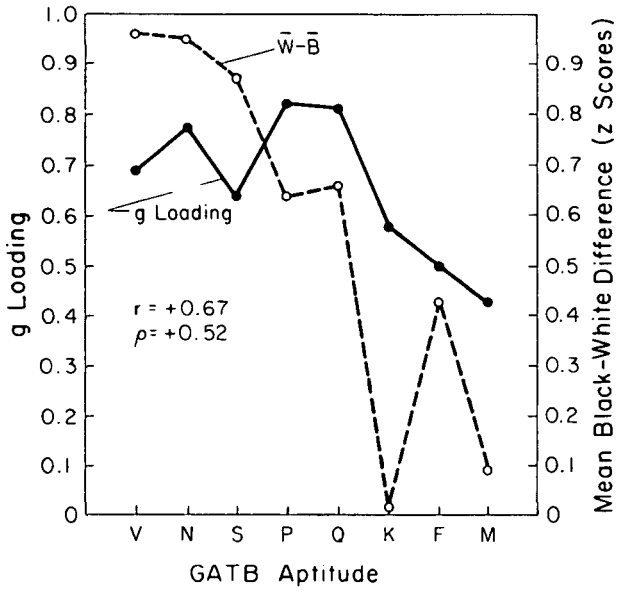


Figure 6. Average differences between black and white samples on the subtests of the General Aptitude Test Battery and the subtests' g loadings. (The differences and the g loadings have been corrected for attenuation.) The correlation between the two profiles is +0.67 (ρ = +0.52). (V—verbal aptitude, N—numerical aptitude, S—spatial aptitude, P—form perception, Q—clerical perception, K—motor coordination, F—finger dexterity, M—manual dexterity.)

the rank-order correlation is 1.00. The Pearson correlation between the *g* loadings and the mean hearing–deaf differences is *negative*, −0.82 (rank-order correlation = −0.70). (The correlation between the profiles of black–white and hearing–deaf mean differences is negative, −0.78.) Obviously, the Spearman hypothesis holds for the black–white differences but not for the differences between hearing and deaf children. Although there are significant ability differences between normally hearing and congenitally deaf children, the differences are not positively related to *g*; if anything, they are *negatively* related to the tests' *g* loadings. The language deprivation caused by deafness evidently takes its toll mostly on common factors other than *g*, or on the tests' specificities (i.e., nonerror variance that is not shared with any other tests in the battery).

Figure 6 shows the same kind of graph for the General Aptitude Test Battery (GATB), with data published by the U.S. Employment Service (U.S. Department of Labor 1970) and based on the test results of more than 27,000 black and white testees. Without correction of *g* loadings and mean differences for attenuation, there is a correlation of +0.71 (ρ = +0.65, $p < .05$) between the *g* loadings and the mean black–white differences, again in accord with Spearman's hypothesis. When the two variates are corrected for attenuation, $r = +0.67$, ρ = +0.52. And so it is for every one of the 11 large data sets I have analyzed. These results are summarized in Table 3. (Each of the data sets is described in the Appendix.) Again, the *g* loadings and the standardized mean black–white differences were corrected for attenuation due to measurement error (test unreliability). These corrections could not be made in four studies for which the tests' reliability coefficients were unavailable, however. The fact that the correlations based on the disattenuated *g* loadings and black–white differences are consistently smaller than the correlations based on the uncorrected *g* loadings and black–white differences (by about 20%) is explained by the considerable decrease in the variability of the loadings and the differences after they are disattentuated. Hence it is this greater restriction of variance on both variables that causes the correlation between them to be more vulnerable to the attenuating effect of sampling error. Another way of examining the effect of test reliability on the Spearman hypothesis is to calculate the partial correlation between *g* loadings (*g*) and mean black–white difference (*d*), with test reliability ($r_{xx}$) partialed out. This was done within each of the seven studies for which reliabilities were available for all of the tests. For these seven sets of data, the average zero-order Pearson correlations between *g* loadings and mean black–white differences are +0.55 for *g* loadings based on the white samples and +0.46 for *g* loadings based on the black samples. The corresponding partial correlations (with test reliabilities partialed out) are +0.53 and +0.36, respectively. (Rank correlations could not be used in this case because partial correlations are not permissible with rank correlation.) The fact that the correlations remain substantial even when the test reliabilities are partialed out contributes further evidence that the Spearman hypothesis is borne out not merely as the result of an artifact of the tests' reliabilities being correlated with both *g* and *d*.

All the evidence reviewed clearly substantiates Spearman's hypothesis (in its weak form). Every set of reasonably suitable data that I have been able to find is consistent with the hypothesis, and I have not been able to find any set of data, based on a diverse collection of tests and on fairly representative samples of the black and white populations, that contradicts the hypothesis.[5] Moreover, no other factors, independent of *g*, extracted in any of these analyses show nearly as large or as consistent correlations with the mean population differences as does the *g* factor.

An important practical implication of Spearman's hypothesis, of course, is that whatever the causes for individual differences and population differences on the general factor of cognitive ability, black people, statistically, will have a greater handicap in those educational, occupational, and military assignments that are most highly correlated with measures of general intelligence. The practical validity of highly *g*-loaded tests for predicting educational and occupational performance and success in the armed forces is the same for the native-born black and white populations in the United States. The practical predictive validity of the *g* of psychometric tests implies that the real-life performance criteria which *g*-loaded tests are capable of predicting with economically consequential accuracy are also *g*-loaded. The practical implications of *g* and Spearman's hypothesis for employment, productivity, and the nation's economic welfare have been discussed in more detail elsewhere (Jensen 1984a).

## Information-processing capacities and psychometric *g*

If the black–white difference is mainly a difference in *g*, then a logical first step toward understanding it scien-

Table 3. *Correlations between g factor loadings and mean black–white differences in 11 test batteries, coefficients of congruence ($r_c$) between black and white g loadings, and percentage of total variance accounted for by the g factor (% Var.) within each population*

| Study[a] | Test | No. of sub-tests | Sample size | | Raw correlation | | | | | | Corrected for attenuation | | | | | | | % Var. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Pearson *r* | | | Spearman rho | | | Pearson *r* | | | Spearman rho | | | | | |
| | | | W | B | W | B | Tot. | W | B | Tot. | W | B | Tot. | W | B | Tot. | $r_c$ | W | B |
| Jensen & Reynolds (1982) | WISC-R | 12 | 1868 | 305 | .73 | .54 | .67 | .73** | .59* | .76** | .71 | .51 | .63 | .59* | .37 | .53* | .995 | 30 | 32 |
| Reynolds & Gutkin (1981)[b] | WISC-R | 12 | 285 | 285 | .51 | .28 | .52 | .56* | .47 | .59* | .47 | .20 | .48 | .36 | .19 | .43 | — | 31 | 33 |
| Sandoval (1982) | WISC-R | 12 | 332 | 314 | .36 | .51 | .50 | .35 | .43 | .47 | .23 | .41 | .34 | .24 | .29 | .26 | .993 | 35 | 33 |
| Mercer (1984) | WISC-R | 12 | 668 | 619 | .66 | .66 | .67 | .71** | .61* | .66* | .59 | .59 | .60 | .42 | .42 | .38 | .998 | 34 | 36 |
| National Longitudinal Study | Various | 12 | 12,275 | 1938 | .78 | .68 | .75 | .43 | .41 | .39 | — | — | — | — | — | — | .995 | 51 | 46 |
| Nichols (1972) | Various | 13 | 1940 | 1460 | .75 | .71 | .74 | .71** | .73** | .67** | — | — | — | — | — | — | .999 | 33 | 34 |
| Dept. of Defense (1982) | ASVAB | 10 | 5533 | 2298 | .39 | .29 | .37 | .30 | .29 | .36 | .31 | .26 | .31 | .15 | .21 | .25 | .995 | 54 | 56 |
| Dept. of Labor (1970) | GATB | 8 | 4001[c] | 2416 | .71 | — | — | .65* | — | — | .67 | — | — | .52 | — | — | — | 35 | — |
| Kaufman & Kaufman (1983) | K-ABC | 13 | 813 | 486 | .56 | .49 | .58 | .59* | .48 | .57* | .53 | .44 | .56 | .54* | .42 | .45 | .997 | 46 | 43 |
| Veroff et al. (1971) | Various | 6 | 179 | 186 | .36 | .32 | .34 | .66 | .60 | .60 | — | — | — | — | — | — | .997 | 53 | 50 |
| Hennessy & Merrifield (1976)[d] | CGP | 10 | 1818 | 431 | .66 | .71 | .70 | .67* | .58* | .54 | — | — | — | — | — | — | .994 | 32 | 38 |
| | | Total: | 29,712 | 10,783 | | | | | | | | | | | | | | | |
| | | | | Mean[e]: | .61 | .54 | .60 | .59 | .53 | .57 | .52 | .41 | .50 | .41 | .32 | .39 | .996 | 39 | 40 |

[a]Study samples, tests, etc. are summarized in the Appendix. [b]Black and white samples matched on socioeconomic status, sex, region of residence, and urban vs. rural residence. [c]*N*s for the black-white difference; *g* loadings from factor analysis of correlations based on a total sample of 27,365 employed workers, high school seniors, college freshmen, basic airmen, and applicants, apprentices, and trainees in various jobs. [d]The data of the black and white groups in this study are statistically adjusted so as to remove the effects of the average population difference in socioeconomic status. [e]All correlations averaged via Fisher's *z* transformation. $^{*}p < .05$ $^{**}p < .01$

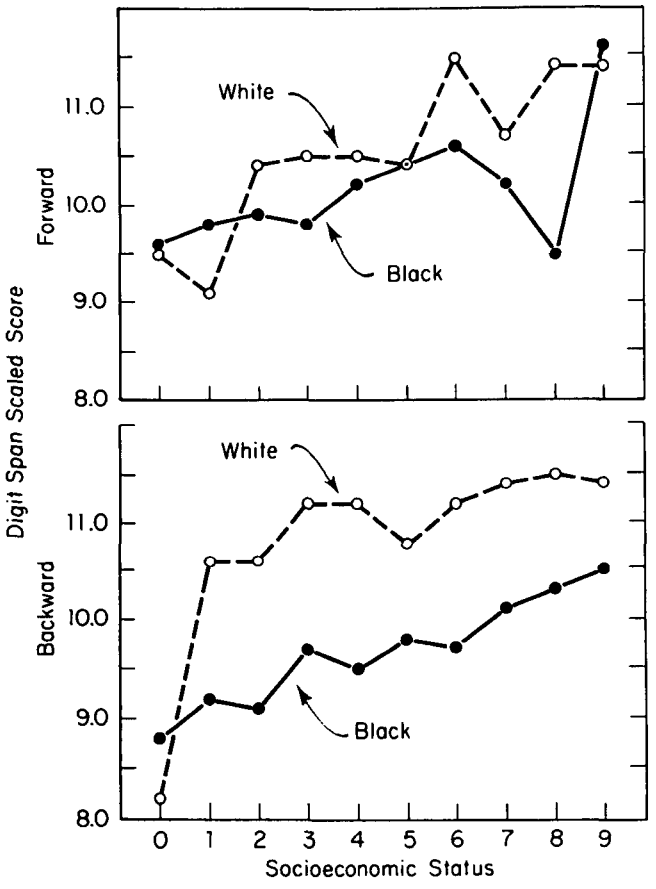


Figure 7. Mean WISC-R forward and backward digit span plotted by socioeconomic level for random samples of 622 black and 622 white school children in California. (From Jensen & Figueroa 1975.)

tifically would be to understand the nature of *g* itself. What is it about a test that makes it more or less *g*-loaded than some other test? Some insight into this question is afforded by the forward and backward digit span subtests of the WISC-R. Neither subtest is a very good measure of *g*, but when they are factor-analyzed among all the other WISC-R subtests, both forward and backward digit span have small to moderate loadings on *g*. But interestingly, as similar as the two tests are in content, they have quite markedly different *g* loadings. Backward digit span is about twice as *g*-loaded as forward digit span (their *g* loadings being close to .50 and .25, respectively), and this is true in both black and white samples. As we might expect, in accord with Spearman's hypothesis, the mean black–white difference is almost twice as great on backward digit span as on forward digit span (Jensen & Figueroa 1975). (This finding was replicated by Jensen and Osborne, 1979.) These results are plotted within each of 10 socioeconomic (SES) categories in Figure 7. (Low SES = 0, high = 9.) On forward digit span, there is a significant main effect for SES, but the population difference and the interaction of population and SES are nonsignificant. In contrast, on backward digit span, both the main effects of SES and population are highly significant ($p < .001$), and, except for the very lowest SES group (SES = 0), there is no significant interaction between population and SES.

How do forward and backward digit span differ in terms of the nature of the cognitive processes involved? The two tasks clearly differ in cognitive complexity. For everyone, backward digits are more difficult than forward digits. Backward span requires more mental manipulation or transformation of the input in order to arrive at the correct output. Presumably, the subject must hold the input series in short-term memory while reversing the order of the digits before "reading" them out. The extra cognitive complexity that this entails, over and above performing the simple forward digit recall, doubles the *g* loading of the task. Hence, in this case *g* seems to reflect the complexity of the mental processes required for a task without being highly related, if related at all, to the specific informational content of the task.

Over the past few years, my graduate students and I have been trying to understand the nature of *g* by means of chronometric analysis of a number of relatively simple tasks that call upon certain elementary cognitive processes but in which there is very little or no intellectual content. All subjects can easily perform the tasks, the only source of reliable individual differences being the speed (measured in milliseconds) with which the subject responds and the degree of consistency in speed of response over a number of trials. On each trial we measure the time it takes for the subject simply to remove his index finger from a push button prior to pressing another button as a means of selecting the choice response. In brief, when the reaction stimulus occurs, the subject removes his finger from the "home" button as quickly as possible and presses one of two (or more) buttons to select the correct response. We measure the time interval between the onset of the reaction stimulus and the removal of the finger from the "home" button. This interval is termed the reaction time (RT). Intraindividual variability (from trial to trial) is measured by the standard deviation of the subject's RT over a number of trials. We have found that RT and intraindividual variability are correlated with IQ and scores on other *g*-loaded tests in children, in the mentally retarded, in university students, and in average adults (Jensen 1979; 1980a; 1980b; 1981; 1982a; 1982b; Jensen & Munro 1979; Jensen, Schafer & Crinella 1981; Sen, Jensen, Sen & Arora 1983; Vernon 1981b; 1983; Vernon & Jensen 1984). Measures of individual differences in choice RT have also shown substantial correlations with scholastic achievement, particularly reading comprehension (with a correlation over .60 in a junior high school sample), even though the RT tasks themselves do not involve reading or any other verbal symbols or scholastic content (Carlson & Jensen 1982). Evidently, certain basic cognitive processes are common to both the RT tasks and scholastic achievement.

In general, more complex RT tasks show higher correlations with IQ or *g* than do simpler tasks (Jensen 1982b). For example, choice RT correlates more highly with IQ than does simple RT; unlike choice RT, simple RT involves no uncertainty and requires no choice or decision. Vernon (1983) did a study in which a battery of RT tests were varied in the types and degree of complexity of their cognitive demands, yet the informational content of the tests was so simple as to be within the capability of most third-grade pupils. The 100 subjects in Vernon's study were university students.

The several tasks and procedures used in Vernon's study are described in detail by Vernon (1983). (The code

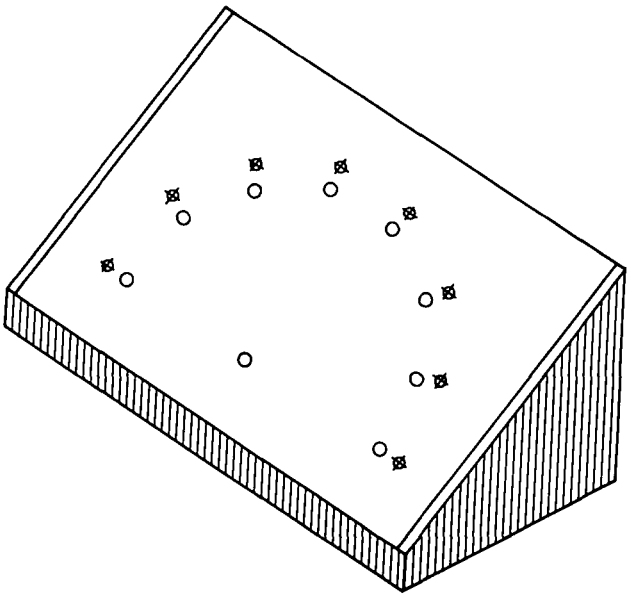

Figure 8. Subject's console of the reaction time aparatus. Pushbuttons indicated by circles, green jewelled lights by circled crosses. The "home" button is in the lower center.

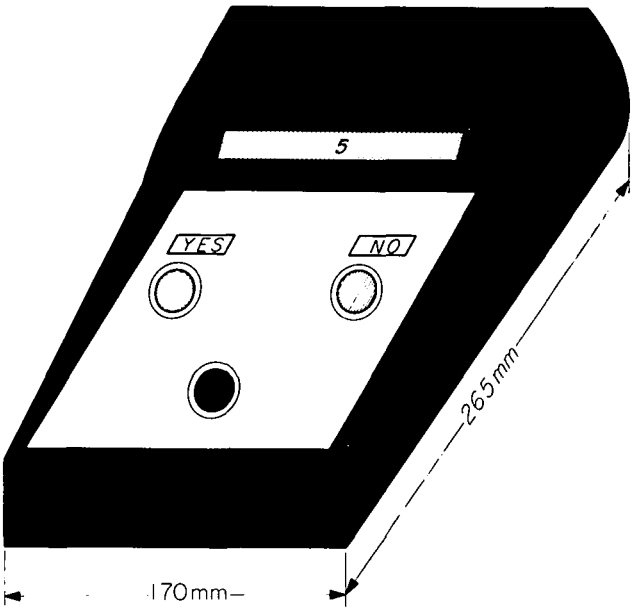


Figure 9. Subject's console used for the digit memory scan, physically same–different words, and synonyms–antonyms test, showing display screen, the two-choice response buttons, and the "home" button (lower center).

symbols by which Vernon labeled each task are given in parentheses in uppercase letters in the following discussion.) Two types of RT apparatus were used. The first is shown in Figure 8. Templates are placed over the console, exposing either 1, 2, 4, or 8 of the light-button combinations. When one of the lights goes on, the subject removes his finger from the central home button and presses a button adjacent to the light, which puts out the light. Fifteen trials are given at each level of complexity – 1, 2, 4, or 8 light–buttons. RT is the time taken to get off the home button after one of the lights goes on. I shall refer to this task simply as the *RT task* (RT). The other tasks all use a two-choice console pictured in Figure 9. In the *Memory Scan* task (DIGIT), a set of digits consisting of anywhere from 1 to 7 digits is simultaneously presented for 2 seconds on the display screen. After a 1-second interval, a single probe digit appears on the screen. The subject's task is to respond as quickly as possible, indicating whether or not the probe was a member of the set that had previously appeared by raising his index finger from the home button and pushing one of the two choice buttons labeled "yes" and "no." The subject's RT is the interval between the onset of the probe digit and the subject's releasing the home button. The subject's score (the average of his RTs to 84 such digit sets) provides a measure of the speed of short-term memory processing, that is, the speed with which information held in short-term memory can be scanned and retrieved.

The *Same–Different* task (SD2) measures the speed of visual discrimination of pairs of simple words that are *physically* the same or different, for example, DOG–DOG or DOG–LOG. The instant that each of 26 pairs of the same or different words is presented, the subject raises his finger from the home button and presses one of the two choice buttons labeled *S* (same) and *D* (different). Again, the subject's RT is the average interval between onset of the word pair and releasing the home button.

The *Synonym–Antonym* task (SA2) works much the same way, but in this test pairs of words are presented that are semantically either similar or opposite in meaning, for example, BIG–LARGE or BIG–LITTLE. All the synonyms and antonyms are composed of extremely common, high-frequency words, and all items can be answered correctly by virtually any third-grader under nonspeeded test conditions. The only reliable source of individual differences is the speed with which the decisions are made. This task measures the subject's speed of access to highly overlearned verbal codes stored in long-term memory.

In the *Dual Processing* tasks, the subject is required to do two things, thus creating some degree of cognitive trade-off, or processing efficiency loss, between storage of information in short-term memory and retrieval of semantic information from long-term memory. In this task, we sequentially combine the digit Memory Scan task and the Same–Different task, or the Memory Scan task and the Synonyms–Antonyms task. First, the subject is presented with a set of 1 to 7 digits for 2 seconds. This presentation is immediately followed by a Same–Different (or Synonym–Antonym) word pair, and the subject must respond "same" or different" (pressing buttons labeled *S* or *D*). Next, the probe digit appears, and he must respond "yes" or "no" to indicate whether or not the probe was a member of the digit set shown previously. The RT (release of home button) is measured for the Same–Different responses to the words (DT2 WORDS) and for the yes–no responses to the probe digits (DT2 DIGITS). The very same dual task procedure is also used with synonyms–antonyms (in place of physically same–different words) and digits (DT3 WORDS and DT3 DIGITS).

As might be expected, the Dual Task, being more complex, elicits slower responses than the combined response times of the component tasks measured separately. It is as if there is some limited central capacity for both working memory and mental processing, and as more of this capacity is used to hold information in memory, less remains available for mental processing of

information. When the task requirements are so complex as to exceed the subject's central capacity, short-term memory and processing break down, and the subject fails to perform the task. Even for less complex tasks, however, in which a complete breakdown does not occur, there is, presumably, some degree of trade-off between storage capacity and processing. The result is a decrement in the efficiency of either, or both, of these functions, and this decrement is reflected in slower reaction time.

When all eight of these information-processing tasks, performed by a group of 100 university students, were factor-analyzed, they yielded a large general factor, accounting for 65.5% of the total variance. This general factor might be termed overall speed of mental processing. It seems a reasonable hypothesis that this is, at least in part, the basis of Spearman's *g*.

This information-processing battery showed a (shrunken) multiple correlation with the Wechsler Adult Intelligence Scale (WAIS) Full Scale IQ of 0.46, which, when corrected for the considerable restriction of range of IQ in this college sample, rose to 0.67. Most noteworthy is the fact that the information-processing battery correlated *only* with the *g* factor of the WAIS. The factor scores derived from the general factor of the reaction times in the information-processing tests and the *g* factor scores of the WAIS are correlated −0.41 ($p < .001$), which would be increased to about −0.60 if corrected for the restriction of range of IQ in the college sample. There is no other shared source of variance, independent of *g*, between the WAIS and the experimental tasks. What is more, the timed subtests of the WAIS showed no higher correlations with the speed of processing measures than did the untimed subtests.

It indeed appears that the WAIS IQ reflects, in part, differences in the speed and efficiency with which individuals can execute a number of elementary cognitive processes. Because the more complex tasks call for more different types of cognitive process, and are also more highly correlated with the *g* of the WAIS, a reasonable hypothesis is that *g* essentially reflects the speed or efficiency with which a number of elementary cognitive processes can be executed. The most highly *g*-loaded tests are those which require the successive or simultaneous execution of a number of these processes. Hence the *g* variance that psychometric tests share with the *g* of RT in information-processing tests does not reflect the specific informational *content* of psychometric tests, but presumably reflects the speed and efficiency of information processing – that is, stimulus encoding, discrimination, comparison, working memory capacity, speed of access and retrieval of information from long-term memory, in addition to certain metaprocesses.

## Elementary cognitive processes in black and white samples

Several independent studies (reviewed in Jensen 1980a, pp. 704–6) have reported significantly greater black–white differences on more complex, choice RT tests than on simple RT. This general finding would seem to be another manifestation of Spearman's hypothesis, as it has also been found that choice RT is more *g*-loaded than simple RT. In order to examine more directly the relationship between RT tests and Spearman's hypothesis, the same battery of eight cognitive processing tasks described in the preceding section was given to 50 black and 56 white male vocational college students, ages 17 to 24 years. (Only those aspects of this study which are most directly germane to Spearman's hypothesis are discussed here; other statistical results are reported elsewhere [Vernon & Jensen 1984].) These subjects were also tested on the Armed Services Vocational Aptitude Battery (ASVAB), a 2½-hour battery that consists of 10 paper-and-pencil tests of typical scholastic knowledge, as well as more specialized knowledge areas: General Science, Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, Numerical Operations, Coding Speed, Auto and Shop Information, and Electronics Information. In a large, representative sample of the nation's population, there is a black–white difference of 1.12 standard deviations on the total ASVAB score (Office of the Assistant Secretary of Defense 1982). Because our vocational college sample was more select and restricted in range of ability than a random sample of the general population, however, the black and white groups of this sample differed by only 0.67σ.

Although the official government publication of the ASVAB survey makes no comment whatsoever regarding the causality of the observed population differences, when the nationwide results on the ASVAB were announced in the general media in 1982, the most common interpretation of the black–white difference was that it could be attributed to the fact that the ASVAB tested mainly scholastic knowledge and skills, and black testees had received generally inferior schooling.

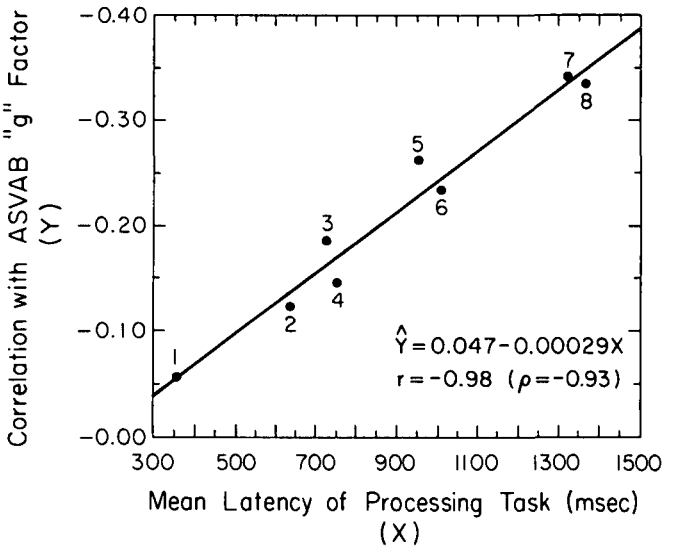


Figure 10. Correlation of processing tasks with ASVAB general factor score as a function of task complexity as indicated by mean response latency (RT in msec) on each task in the total vocational college sample ($N$ = 106). (The numbers beside the data points indicate the specific processing tasks: 1—RT, 2—DIGIT, 3—DT2 Digits, 4—DT3 Digits, 5—SD2, 6—DT2 Words, 7—DT3 Words, 8—SA2.) (From Vernon & Jensen 1984.) [*Editorial note:* In the version of the target article seen by the commentators Figure 10 contained some technical errors which were subsequently drawn to the author's attention by L. V. Jones in his commentary (q.v.). The corrections are made here but they are drawn to the reader's attention because of BBS's policy that no substantive changes can be made after the commentators have seen the preprint.]

Our mental processing battery of reaction time tests obviously differs markedly from the ASVAB in terms of scholastic or intellectual content. Yet the processing battery showed a significant ($p < .01$) (shrunken) multiple correlation of about 0.5 with the total ASVAB score, in both the black and the white samples. The degree of correlation between the processing tests and the ASVAB, moreover, was directly related to the complexity of the processing tests. When the mean response latency on each processing test was used as the only available objective index of complexity, the correlation between the profile of these mean latencies on each of the eight tasks and the profile of the correlations of each task with the general factor score of the ASVAB was $r = -0.98$, ($\rho = -0.93$, $p < .01$), as shown in Figure 10 [*see editorial note in figure caption*]. In other words, *the more complex the processing required by the different cognitive processing tasks, the stronger was their relationship to the g factor of the ASVAB.* The correlation between the mental processing tests and the ASVAB cannot be attributed to the fact that two of the ASVAB tests are speeded. Indeed, the most speed-dependent subtest in the ASVAB, the Coding Speed test, proved to be the *least* correlated with the processing tests, and also showed the lowest *g* loading among all 10 subtests of the ASVAB.

A discriminant function analysis was performed using the 10 ASVAB tests to determine the maximal discrimination this combination of tests could make between the black and the white samples. A single discriminant function correctly classified 73% of the subjects as black or white. (This result can also be expressed as a multiple correlation of 0.51 [shrunken = 0.42] between the ASVAB and the black–white classification.)

A discriminant analysis was also applied to all the variables yielded by our battery of mental processing tests. A single discriminant function correctly classified 72% of the subjects as black or white (multiple $R = 0.52$, shrunken = 0.37, $p < .01$). The black–white differences on the separate speed-of-processing variables, however, were very small and generally nonsignificant for any single variable. They discriminate significantly between the populations only when analyzed all together in combination, as seen in the discriminant function analysis, in large part because the relative magnitudes of the differences on the various processing tasks are in close accord with Spearman's hypothesis and the idea that *g* reflects cognitive complexity. As shown in Figure 11, there is a correlation between the magnitudes of the mean black–white differences in response latency on the eight mental processing variables and the variables' cognitive complexity as objectively indexed by their mean latencies in the combined samples.

As seen in Figure 12, an even stronger relationship between task complexity and group differences in response latency was found when vocational college students ($N = 106$) and university students ($N = 100$) were compared. These groups differ more markedly in psychometric *g* than do the black and white vocational college samples in the present study. (Details of these comparisons are given by Vernon & Jensen, 1984.)

When we test Spearman's hypothesis with this mental processing battery in the same fashion as we have previously tested Spearman's hypothesis in all the other test batteries, the Pearson *r* between the eight processing variables' *g* loadings and the corresponding standardized mean black–white differences is +0.40, $\rho = +0.38$. (A test of Spearman's hypothesis based on the ASVAB in the present black and white samples shows a Pearson correlation of +0.59 [$\rho = +0.37$] between the *g* loadings [with

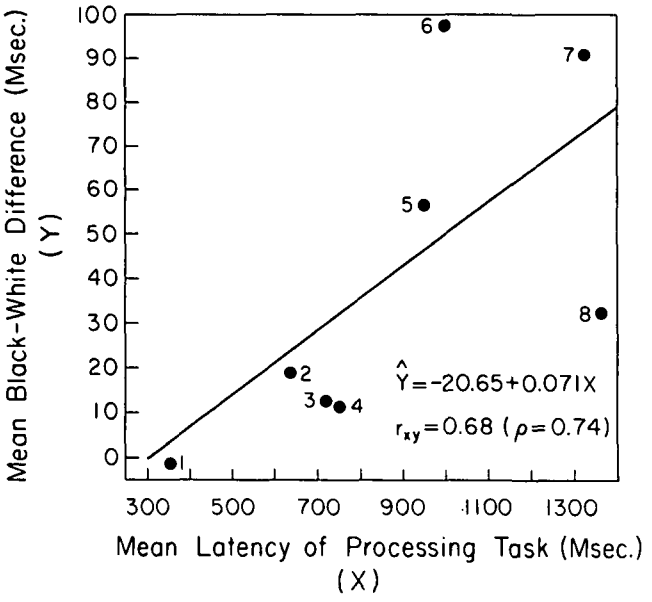


Figure 11. Mean black–white difference (in msec) in response latency (RT) to various processing tasks as a function of task complexity as indicated by mean response latency (in msec) on each task in the combined groups. (Tasks are identified by the numbered data points, as listed in the caption of Figure 10.) (From Vernon & Jensen 1984.)

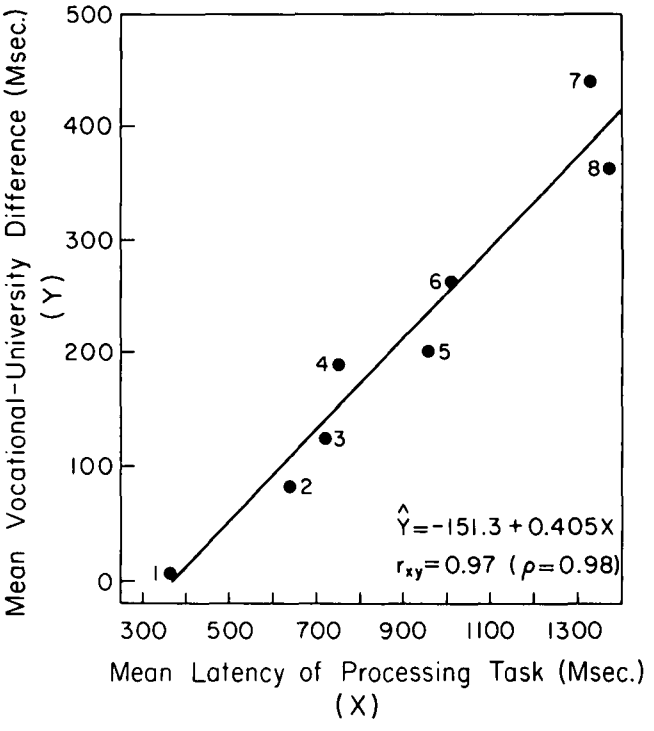


Figure 12. Mean difference (in msec) between vocational college students ($N = 106$) and university students ($N = 100$) on various processing tasks as a function of task complexity as indicated by mean response latency (RT) on each of the tasks in the vocational college group. (The tasks are identified by the numbered data points, as listed in the caption of Figure 10.) (From Vernon & Jensen 1984.)

variance attributable to black–white differences partialed out] of the 10 ASVAB subtests and the standardized mean black–white differences on the subtests.)

One manifestation of even small differences in rates of information processing that would be expected over an extended period of time (the years of formal education, for example) is a difference in the acquisition of knowledge and problem-solving skills required on standard tests of mental ability and achievement tests such as the ASVAB. Some considerable part of the variance in achievement-oriented psychometric tests like the ASVAB reflects what Cattell (1963) has termed "crystalized" intelligence. Because measures of elementary information processing seem to be closer to tests of "fluid" intelligence, such as Raven's Matrices, one might expect a higher correlation between speed of mental processing and the *g* factor of tests of "fluid" intelligence. As the present black–white difference in general speed of processing is only about one-third as large as the mean black–white difference on the ASVAB, it seems likely that the *g* of the ASVAB (and similar achievement-oriented psychometric tests) also involves types of higher-order processing other than the quite elementary processes measured by the present tasks, namely, the metaprocesses that control the deployment of the elementary processes in the solution of complex problems (Sternberg & Gardner 1982). A potentially fruitful conjecture is that a large part of the black–white difference on the ASVAB may be attributable to differences in certain metaprocesses as well as to differences in the elementary processes. Exploration of such hypotheses is the task of future research.

## The future of this line of research

I believe that Spearman's hypothesis has been substantiated in psychometric test data, and that we have made a good beginning to investigating its possible locus in the speed or efficiency of various cognitive processes, as measured by reaction-time techniques. But the processes we have succeeded in measuring thus far account neither for the whole of psychometric *g* nor for more than a small fraction of the total black–white difference on typical psychometric tests. And *g* itself, although a major source of variance, does not represent the totality of mental abilities. We know there are a good many other ability factors besides *g*, even though their relative importance in a technological society such as ours may be overshadowed by the pervasive educational and occupational demands on *g*, as I have spelled out elsewhere (Jensen 1984a).

A useful analogy may be drawn between cognitive processes and computers, likening some processes to the hardware and some to the software components. In terms of this analogy, we are still very uncertain about the relative degrees to which psychometric *g* reflects the "hardware" and "software" components of cognitive processing. Yet it is essential that we learn more if we are to direct our educational efforts most productively. It seems likely that the "software" components of intelligent behavior (the so-called metaprocesses of executive control, problem-solving strategies, predicting and monitoring one's own performance, and the like) may be more readily trainable than the "hardware" components (speed of encoding, short-term memory capacity, retrieval of information in long-term memory, etc.), which are presumably more closely linked to the neural substrate of mental activity. We are even uncertain to what extent these hardware components of human information processing are amenable to special training (Detterman & Sternberg 1982; Jensen 1983c).

By investigating these kinds of questions with the types of reaction-time techniques capable of measuring a variety of elementary cognitive processes and metaprocesses, we can hope to make further progress toward understanding the nature of *g* and, ultimately, toward understanding precisely the nature of the processes underlying various individual and group differences in human mental ability.

## APPENDIX

### Notes on the 11 studies used in the analysis of Spearman's hypothesis

The tests used in the 11 studies that provided data suitable for testing Spearman's hypothesis are listed (with code numbers) in Table 4. There are 74 distinct tests in all, but many are quite similar in the types of content and skills they include. For example, there are five different vocabulary tests and six different tests of arithmetic reasoning or computation.

Table 5 gives the sample sizes in each study, and the tests in each study are listed by their code numbers (from Table 4), along with the ($\bar{D}$) mean black–white difference in $\sigma$ units, ($g$) the *g* loadings obtained separately (when possible) within the black and white samples, and ($r_{xx}$) the test's reliability coefficient (when available) for the study sample or a closely comparable sample.

For details of each study, readers are referred to the cited sources. The most essential information with respect to the present analyses is summarized below.

#### *Jensen and Reynolds (1982)*

**SAMPLE.** This study used a national standardization sample for the Wechsler Intelligence Scale for Children–Revised, selected by a stratified random sampling procedure to be representative of the entire U.S. population based on the 1970 census. Black and white age-matched samples of children between the ages of 6 and 16½ years were tested.

**ANALYSIS.** A Schmid–Leiman hierarchical factor analysis (Schmid & Leiman 1957) of the 13 subtests was done separately in the black and white samples, all ages combined. Because all scores are age-standardized, age variance does not enter into the test intercorrelations or factor loadings. The congruence coefficients between the Schmid–Leiman *g*, the first principal factor, and the first principal component are all +0.999 in both populations.

To get some idea of the effects of sampling error on the mean black–white differences ($\bar{D}$) and the *g* loadings, the two samples were each split randomly in half, and $\bar{D}$ and *g* loadings of each test were determined within the random halves. The Pearson correlations between the profiles of $\bar{D}$ and of *g* loadings were corrected for attenuation using these Spearman-Brown boosted split-half reliabilities, yielding corrected correlations of +0.62 and +0.81 for the black and white samples, respectively. (Further details of this analysis are given in Jensen & Reynolds, 1982, pp. 433–35.)

Three other factors (all orthogonal to *g* and to each other) emerge in the Schmid–Leiman analysis: Verbal (V), Performance (P), and Memory (M). The correlations of the standard-

Table 4. *Master list of the tests used in 11 studies*

| Code no. | Test | Study[a] |
|---|---|---|
| 1 | WISC-R[b]: Information | J R S M N |
| 2 | WISC-R: Similarities | J R S M |
| 3 | WISC-R: Arithmetic | J R S M |
| 4 | WISC-R: Vocabulary | J R S M N |
| 5 | WISC-R: Comprehension | J R S M N |
| 6 | WISC-R: Digit Span | J R S M N |
| 7 | WISC-R: Picture Completion | J R S M |
| 8 | WISC-R: Picture Arrangement | J R S M N |
| 9 | WISC-R: Block Design | J R S M N |
| 10 | WISC-R: Object Assembly | J R S M |
| 11 | WISC-R: Coding | J R S M N |
| 12 | WISC-R: Mazes | J R S M |
| 13 | WISC-R: Tapping Span (Knox Cubes) | J |
| 14 | CGP[c]: Vocabulary | NL, H |
| 15 | CGP: Picture-Number (Paired-Associates Memory) | NL, H |
| 16 | CGP: Reading | NL, H |
| 17 | CGP: Letter Groups (Inductive Reasoning) | NL, H |
| 18 | CGP: Math | NL |
| 19 | CGP: Mosaic Comparisons (Perceptual Speed and Accuracy) | NL, H |
| 20 | SAT[d]–Verbal | NL |
| 21 | SAT–Math | NL |
| 22 | ACT[e]–English | NL |
| 23 | ACT–Social Studies | NL |
| 24 | ACT–Science Reading Comprehension | NL |
| 25 | ACT–Math | NL |
| 26 | Bender-Gestalt (Form Perception) | N |
| 27 | ITPA[f]: Auditory-Vocal Association | N |
| 28 | Draw-A-Man | N |
| 29 | WRAT[g]: Spelling | N |
| 30 | WRAT: Reading | N |
| 31 | WRAT: Arithmetic | N |
| 32 | ASVAB[h]: General Science | D |
| 33 | ASVAB: Arithmetic Reasoning | D |
| 34 | ASVAB: Word Knowledge (Vocabulary) | D |
| 35 | ASVAB: Paragraph Comprehension | D |
| 36 | ASVAB: Numerical Operations (Computation) | D |
| 37 | ASVAB: Coding Speed | D |
| 38 | ASVAB: Auto-Shop Information | D |
| 39 | ASVAB: Mathematics Knowledge | D |
| 40 | ASVAB: Mechanical Comprehension | D |
| 41 | ASVAB: Electronics Information | D |
| 42 | GATB[i]: V–Verbal Aptitude (Vocabulary) | L |
| 43 | GATB: N–Numerical (Computation and Arithmetic Reasoning) | L |
| 44 | GATB: S–Spatial (3-Dimensional Space) | L |
| 45 | GATB: P–Form Perception (Tool Matching and Form Matching) | L |
| 46 | GATB: Q–Clerical Perception (Name Comparison) | L |
| 47 | GATB: K–Motor Coordination (Mark Making) | L |
| 48 | GATB: F–Finger Dexterity (Assemble and Disassemble) | L |
| 49 | GATB: M–Manual Dexterity (Place and Turn) | L |
| 50 | K-ABC[j]: Hand Movements | K |
| 51 | K-ABC: Number Recall | K |
| 52 | K-ABC: Word Order | K |
| 53 | K-ABC: Gestalt Closure | K |
| 54 | K-ABC: Triangles | K |
| 55 | K-ABC: Matrix Analogies | K |
| 56 | K-ABC: Spatial Memory | K |
| 57 | K-ABC: Photo Series | K |
| 58 | K-ABC: Faces and Places | K |
| 59 | K-ABC: Arithmetic | K |
| 60 | K-ABC: Riddles | K |
| 61 | K-ABC: Reading (Decoding) | K |
| 62 | K-ABC: Reading (Comprehension) | K |
| 63 | WAIS[k]: Digit Span | V |
| 64 | Lorge–Thorndike: Sentence Completion | V |
| 65 | Raven Progressive Matrices | V |
| 66 | Ammons Quick Test (Picture Vocabulary) | V |
| 67 | WAIS[k]: Information | V |
| 68 | WAIS: Coding (Digit Symbol Substitution) | V |
| 69 | CGP[c]: Sentences (Grammatical Usage) | H |
| 70 | CGP: Year 2000 (Integrative Reasoning) | H |
| 71 | CGP: Intersections (Spatial Reasoning) | H |
| 72 | CGP: Information About Technology | H |
| 73 | CGP: Algebra | H |

[a]Studies indicated by the following letter codes: J: Jensen & Reynolds (1982); R: Reynolds & Gutkin (1981); S: Sandoval (1982); M: Mercer (1984); NL: National Longitudinal Study; N: Nichols (1972); D: Department of Defense (1982); L: Department of Labor (1970); K: Kaufman & Kaufman (1983); V: Veroff et al. (1971); H: Hennessy & Merrifield (1976). [b]Wechsler Intelligence Scale for Children–Revised. [c]Comparative Guidance and Placement Program. [d]Scholastic Aptitude Test. [e]American College Test. [f]Illinois Test of Psycholinguistic Abilities. [g]Wide-Range Achievement Test. [h]Armed Services Vocational Aptitude Battery. [i]General Aptitude Test Battery. [j]Kaufman Assessment Battery for Children. [k]Wechsler Adult Intelligence Scale.

(*continued*)

Table 5. *Mean black–white difference, $\bar{D}$ (in σ units), g loading (for black sample [B] and white sample [W]), and reliability ($r_{xx}$) of tests in 11 studies ($\bar{D}$, g, $r_{xx}$, all × 100)*

| Study | Test code | $\bar{D}$ | $g_W$ | $g_B$ | $r_{xx}$ |
|---|---|---|---|---|---|
| Jensen & Reynolds | 1 | 81 | 67 | 65 | 85 |
| Black $N$ = 305 | 2 | 79 | 67 | 62 | 81 |
| White $N$ = 1868 | 3 | 61 | 57 | 60 | 77 |
| | 4 | 88 | 72 | 71 | 86 |
| | 5 | 94 | 60 | 61 | 77 |
| | 6 | 31 | 44 | 59 | 78 |
| | 7 | 79 | 51 | 57 | 77 |
| | 8 | 77 | 49 | 49 | 73 |
| | 9 | 93 | 65 | 61 | 85 |
| | 10 | 82 | 50 | 53 | 70 |
| | 11 | 47 | 37 | 36 | 72 |
| | 12 | 69 | 37 | 45 | 72 |
| | 13 | 33 | 35 | 44 | 80 |
| Reynolds & Gutkin | 1 | 69 | 67 | 65 | 85 |
| Black $N$ = 285 | 2 | 53 | 67 | 62 | 81 |
| White $N$ = 285 | 3 | 48 | 57 | 60 | 77 |
| | 4 | 67 | 72 | 71 | 86 |
| | 5 | 80 | 60 | 61 | 77 |
| | 6 | 12 | 44 | 59 | 78 |
| | 7 | 61 | 51 | 57 | 77 |
| | 8 | 65 | 49 | 49 | 73 |
| | 9 | 73 | 65 | 61 | 85 |
| | 10 | 64 | 50 | 53 | 70 |
| | 11 | 39 | 37 | 36 | 72 |
| | 12 | 59 | 37 | 45 | 72 |
| Sandoval | 1 | 93 | 73 | 71 | 85 |
| Black $N$ = 314 | 2 | 84 | 68 | 60 | 81 |
| White $N$ = 332 | 3 | 63 | 69 | 65 | 77 |
| | 4 | 78 | 73 | 70 | 86 |
| | 5 | 65 | 64 | 65 | 77 |
| | 6 | 49 | 58 | 41 | 78 |
| | 7 | 63 | 49 | 56 | 77 |
| | 8 | 76 | 55 | 60 | 73 |
| | 9 | 96 | 61 | 61 | 85 |
| | 10 | 81 | 58 | 52 | 70 |
| | 11 | 50 | 33 | 40 | 72 |
| | 12 | 83 | 34 | 42 | 72 |
| Mercer | 1 | 101 | 70 | 71 | 85 |
| Black $N$ = 619 | 2 | 82 | 67 | 60 | 81 |
| White $N$ = 668 | 3 | 68 | 65 | 66 | 77 |
| | 4 | 90 | 72 | 74 | 86 |
| | 5 | 65 | 63 | 64 | 77 |
| | 6 | 52 | 49 | 48 | 78 |
| | 7 | 65 | 50 | 58 | 77 |
| | 8 | 79 | 54 | 61 | 73 |
| | 9 | 89 | 59 | 62 | 85 |
| | 10 | 82 | 58 | 58 | 70 |
| | 11 | 46 | 35 | 42 | 72 |
| | 12 | 81 | 42 | 45 | 72 |
| National Longitudinal Study | 14 | 100 | 73 | 65 | 90 |
| Black $N$ = 1,938 | 15 | 65 | 39 | 38 | |
| White $N$ = 12,275 | 16 | 99 | 77 | 74 | 81 |
| | 17 | 105 | 64 | 65 | 75 |
| | 18 | 109 | 81 | 75 | |
| | 19 | 92 | 34 | 44 | 77 |
| | 20 | 115 | 87 | 91 | |
| | 21 | 121 | 80 | 78 | |
| | 22 | 116 | 73 | 71 | |
| | 23 | 120 | 76 | 74 | |
| | 24 | 124 | 75 | 61 | |
| | 25 | 118 | 75 | 59 | |
| Nichols | 26 | 69 | 56 | 59 | |
| Black $N$ = 1460 | 1 | 37 | 58 | 63 | 66 |
| White $N$ = 1940 | 5 | 41 | 44 | 43 | 59 |
| | 4 | 85 | 61 | 63 | 77 |
| | 6 | 45 | 57 | 54 | 60 |
| | 8 | 71 | 58 | 59 | 72 |
| | 9 | 66 | 51 | 55 | 84 |
| | 11 | 17 | 31 | 25 | 60 |
| | 27 | 96 | 69 | 66 | |
| | 28 | 11 | 44 | 49 | |
| | 29 | 73 | 69 | 70 | |
| | 30 | 73 | 67 | 69 | |
| | 31 | 55 | 71 | 69 | |
| | 32 | 123 | 84 | 83 | 86 |
| Black $N$ = 2298 | 33 | 116 | 85 | 76 | 87 |
| White $N$ = 5533 | 34 | 130 | 82 | 87 | 86 |
| | 35 | 108 | 73 | 81 | 68 |
| | 36 | 95 | 62 | 71 | 71 |
| Dept. of Defense | 37 | 96 | 51 | 63 | 82 |
| | 38 | 123 | 59 | 65 | 83 |
| | 39 | 88 | 80 | 77 | 84 |
| | 40 | 120 | 74 | 66 | 83 |
| | 41 | 122 | 77 | 74 | 80 |
| Dept. of Labor | 42 | 89 | 64 | | 86 |
| Black $N$ = 2416 | 43 | 87 | 71 | | 84 |
| White $N$ = 4401 | 44 | 78 | 58 | | 81 |
| | 45 | 55 | 70 | | 73 |
| | 46 | 57 | 71 | | 75 |
| | 47 | 2 | 52 | | 81 |
| | 48 | 35 | 41 | | 67 |
| | 49 | 8 | 37 | | 73 |
| Kaufman | 50 | 57 | 56 | 50 | 76 |
| Black $N$ = 486 | 51 | 4 | 59 | 51 | 81 |
| White $N$ = 813 | 52 | 15 | 62 | 66 | 82 |
| | 53 | 39 | 44 | 50 | 71 |
| | 54 | 61 | 67 | 55 | 84 |
| | 55 | 48 | 67 | 61 | 85 |
| | 56 | 47 | 58 | 54 | 80 |
| | 57 | 56 | 64 | 61 | 82 |
| | 58 | 38 | 73 | 74 | 84 |
| | 59 | 82 | 81 | 78 | 87 |
| | 60 | 88 | 80 | 79 | 86 |

(*continued*)

(*continued*)

Table 5 (*Continued*)

| Study | Test code | $\bar{D}$ | $g_W$ | $g_B$ | $r_{xx}$ |
|---|---|---|---|---|---|
| Kaufman | 61 | 49 | 80 | 77 | 92 |
| | 62 | 65 | 81 | 79 | 91 |
| Veroff et al. | 63 | 41 | 68 | 55 | 68 |
| Black $N$ = 186 | 64 | 55 | 81 | 83 | |
| White $N$ = 179 | 65 | 103 | 76 | 69 | |
| | 66 | 81 | 73 | 78 | |
| | 67 | 43 | 69 | 66 | 91 |
| | 68 | 79 | 71 | 68 | 92 |
| Hennessy | 16 | 65 | 75 | 79 | 81 |
| Black $N$ = 431 | 69 | 78 | 68 | 68 | 84 |
| White $N$ = 1818 | 70 | 84 | 71 | 74 | 73 |
| | 19 | 71 | 40 | 58 | 77 |
| | 17 | 69 | 60 | 70 | 75 |
| | 14 | 52 | 61 | 68 | 90 |
| | 15 | 60 | 34 | 35 | |
| | 71 | 19 | 31 | 25 | |
| | 72 | 41 | 43 | 55 | |
| | 73 | 79 | 62 | 64 | 88 |

ized black–white differences with the loadings of the 13 subtests on each of these factors in each population are as follows:

| | *Black* | *White* |
|---|---|---|
| V | +0.49 | +0.49 |
| P | +0.45 | +0.39 |
| M | −0.88 | −0.89 |

Note that the black subjects are superior to the white on the short-term memory factor (independent of g, V, and P).

### *Reynolds and Gutkin (1981)*

**SAMPLES.** The white subjects in this study are a subset of the total white sample ($N$ = 1870) used in the WISC-R national standardization. All of the black subjects ($N$ = 305) in the national standardization were considered for this study. (Subjects in both samples ranged in age from 6 to 16½ years. The WISC-R scores were age-standardized.) It was possible to obtain exact matches of 285 black testees with 285 white testees on four demographic variables (sex, SES, geographic region of residence, and urban vs. rural residence). The mean black–white differences on the WISC-R subtests are based on this sample of 285 matched pairs. The g loadings (Schmid–Leiman) were based on the total black ($N$ = 305) and white ($N$ = 1868) standardization samples. The results show that matching black and white samples on SES (and the other variables) reduces the overall IQ difference between the groups but has little effect on the profile of subtest differences. The SES (and demographically) matched black and white samples differ by 12.34 IQ points (or 0.92 σ units) on Full Scale IQ as compared with the unmatched total national standardization samples, which show a mean black–white difference on Full Scale IQ of 15.83 IQ points (or 1.14 σ units). The profile of 12 black–white mean differences based on the entire standardization sample is correlated +0.97 with the profile of 12 black–white mean differences based on the demographically matched subsets of 285 black and 285 white children.

### *Sandoval (1982)*

**SAMPLES.** Ss are a subsample of the children used to standardize the System of Multiculture Pluralistic Assessment (SOMPA), which includes the WISC-R. (This sample is independent of the WISC-R national standardization sample.) The total sample was selected by a random-stratified (by sex, ethnicity, age, locality) sampling procedure to be representative of the California elementary-school-age population (5 to 11 years).

**ANALYSIS.** Principal factor analysis was applied to the data. This study also included Mexican-American children, who were not included in the present analysis, because no prediction was made for this group with respect to Spearman's hypothesis. However, Sandoval also examined Spearman's hypothesis on the white-Anglo versus Mexican-American groups and found a rank-order correlation of +0.78 between the Anglo/Mexican-American differences on the 12 WISC-R subtests and their g loadings. The profiles of Anglo/black subtest differences and Anglo/Mexican-American subtest differences are correlated only +0.29. However, the WISC-R factors (g, V, P) are almost identical in the Anglo, black, and Mexican-American samples. (Congruence coefficients for the g loadings between the three populations range between .99 and 1.00.)

### *Mercer (1984)*

**SAMPLES.** These are all the black and white subjects in the SOMPA standardization sample, randomly selected from all 5- to 11-year-olds in the California school population in 1973–74.

**ANALYSIS.** Principal factor analysis was performed on the data. Mercer also provides the correlations (corrected for contamination) between each of the 12 WISC-R subscales and Full Scale IQ. If Full Scale IQ is a rough estimate of the general factor of the WISC-R, it is interesting to note the degree of relation between the profile of the subtests' correlations with Full Scale IQ and the profile of the subtests' g factor (i.e., first principal factor) loadings. The correlations between these two profiles are +0.91 within the black sample and +0.91 within the white sample.

### *National Longitudinal Study*

The National Longitudinal Study of Educational Effects (NLS), conducted by the National Center for Educational Statistics (NCES), was based on a large, stratified-random sample of U.S. high school graduates of 1972. The data tapes may be purchased from the NCES, Department of Health Education, and Welfare, Washington, D.C. Principal factor analysis was used.

### *Nichols (1972)*

**SAMPLES.** This doctoral study by Nichols provides relevant data on seven subscales of the Wechsler Intelligence Scale for Children (WISC) in addition to six other cognitive ability and achievement tests given to large black ($N$ = 1460) and white ($N$ = 1940) samples of seven-year-old children in several large cities in the United States. All were participants in a large-scale longitudinal study (the Collaborative Study) conducted by the National Institutes of Health. The subjects were enlisted in 12 public hospitals at the time of their mothers' pregnancy, and they are a fairly representative sample of the populations served by these large city hospitals, a population that Nichols describes as "skewed somewhat to the lower end" in social class.

**ANALYSIS.** The correlation matrices (for black and white subjects separately) were subjected to a Schmid–Leiman hierarchical factor analysis. The coefficient of congruence between the hierarchical g factor and the first principal factor is +0.999 in

both black and white samples, and the congruence coefficient between the *g* factor of black and white samples for both types of factor analysis is +0.999.

### *Department of Defense (1982)*

**SAMPLE.** This is described in detail in *Profile of American Youth* (March 1982). This study went to great pains to obtain a large nationwide probability sample representative of the population of American youths of ages 16 to 23 years.

**ANALYSIS.** Because the official publication on this study presents only the intercorrelations between the 10 subtests of the Armed Services Vocational Aptitude Battery (ASVAB) for the total national sample, the correlation matrices for black and white samples separately were obtained from the Office of the Assistant Secretary of Defense in 1982, after the ASVAB survey data were declared in the public domain. The correlation matrices, separately for black and white samples, were subjected to a principal factor analysis. Also, both samples were randomly split in half to determine the sampling reliability of the profiles of black–white differences and *g* loadings. All of these reliabilities in both populations are so high (ranging from .98 to .999) that correction for attenuation of the correlation between the subtests' profile of black–white differences and the profile of *g* loadings would have virtually no effect.

One may wonder, then, why the ASVAB, although it yielded the largest *g* loadings and the largest black–white differences of any of the 11 test batteries in the present study, shows the lowest correlation between the profiles of black–white differences and *g* loadings. The answer appears to be that the ASVAB subtests are all so highly *g*-loaded, with so very little variation in their *g* loadings, that the effect of other factors and specificity in the subtests dominates the variation in the profile of black–white differences on the 10 subtests, even though the reliable non-*g* factors (and specificities) constitute only a relatively small proportion (about .20) of the total variance in the ASVAB subtest scores. Because of this, the ASVAB is probably the least ideal battery for testing Spearman's hypothesis. Cronbach (1979) has questioned the use of the ASVAB in educational and vocational counseling, essentially because the rather uniformly high *g* loadings of all of the subtests leave too little non-*g* variance to obtain sufficiently reliable or predictively valid differential patterns of the subtest scores for individuals.

Reliabilities of the ASVAB subtests are provided by Bock and Mislevy (1981, Tables 2 and 3).

### *Department of Labor (1970)*

**SAMPLES.** In its *Manual* for the General Aptitude Test Battery (GATB), the U.S. Employment Service (Manpower Administration, U.S. Department of Labor) gives the intercorrelations of the nine GATB aptitudes (Tables 6-5 through 6-9, pp. 32–34). The correlations were not computed separately for black and white samples but are based on predominantly white samples. The correlations are based on very large samples (total $N$ = 27,365) of employed workers, high school seniors, college freshmen, basic airmen, and applicants, apprentices, and trainees in various jobs. The *g* loadings of the GATB aptitudes in these five samples are so highly similar as to justify averaging them over the five samples. Using analysis of variance, the profile of these averaged *g* loadings on the GATB aptitudes has a profile reliability of 0.96. The reliabilities of each of the GATB aptitudes are given in the *Manual* (p. 255).

The mean black–white differences (in σ units) were obtained from separate reports put out by the USES; each report gives means and standard deviations of the GATB aptitudes for black and white subjects in various occupations. When the present

Table 6. *Mean black–white differences (in σ units) on GATB aptitudes (determined from data on 33 occupational samples provided in the following Technical Reports of the United States Employment service of the U.S. Department of Labor Manpower Administration)*

| Report No. | *N* W | *N* B | Date | Title |
|---|---|---|---|---|
| S-447 | 59 | 57 | 1969 | Production-line welder |
| S-465 | 34 | 31 | 1972 | Covering machine operator |
| S-343R | 224 | 46 | 1973 | Operating engineer (construction work) |
| S-131R74 | 95 | 91 | 1974 | Fork-lift truck operator |
| S-180R74 | 205 | 120 | 1974 | Keypunch operator |
| S-239R74 | 99 | 81 | 1974 | Medical ward clerk |
| S-266R74 | 221 | 40 | 1974 | Drafter |
| S-310R74 | 103 | 59 | 1974 | Electronics assembler |
| S-329R74 | 225 | 130 | 1974 | General office clerk |
| S-398R74 | 161 | 91 | 1974 | Teacher aide (elementary school) |
| S-217R75 | 127 | 61 | 1975 | Banking proof-machine operator |
| S-228R75 | 72 | 67 | 1975 | Injection-molding machine tender |
| S-259R75 | 168 | 78 | 1975 | Bank teller |
| S-270R75 | 118 | 73 | 1975 | Practical nurse |
| S-282R75 | 68 | 66 | 1975 | Nurse aide |
| S-370R75 | 111 | 30 | 1975 | Production and maintenance mechanic |
| S-115R76 | 106 | 49 | 1976 | Weaver (carpet & rug, textile) |
| S-135R76 | 126 | 83 | 1976 | Production-machine operator |
| S-144R76 | 57 | 30 | 1976 | Woodworking-machine operator |
| S-145R76 | 42 | 42 | 1976 | Grocery checker |
| S-335R76 | 70 | 50 | 1976 | Extruding-machine (wire) operator |
| S-381R76 | 43 | 44 | 1976 | Electronics micrologic assembler |
| S-74R77 | 102 | 39 | 1977 | Telephone repairer |
| S-101R77 | 138 | 57 | 1977 | Automobile assembler |
| S-276R77 | 110 | 57 | 1977 | Salesperson, general merchandise |
| S-309R77 | 97 | 63 | 1977 | Banking encoder |
| S-414R77 | 115 | 56 | 1977 | Electrical equipment assembler |
| S-61R78 | 184 | 46 | 1978 | Plumber, pipe fitter |
| S-327R78 | 123 | 129 | 1978 | Psychiatric technician |
| S-467R78 | 141 | 109 | 1978 | Capacitor assembler |
| S-468R78 | 41 | 21 | 1978 | Cigarette inspector |
| S-469R78 | 155 | 78 | 1978 | Chemical operator |
| S-471R81 | 219 | 321 | 1981 | Semiconductor occupations |

study was completed, 33 such reports were available from the USES; these are listed in Table 6. The standardized mean black–white differences on each of the eight GATB aptitudes were averaged over all 33 occupational groups. Analysis of variance was used to determine the reliability of the profile of averaged black–white differences on the eight aptitudes; the profile reliability is 0.97.

The matrix of correlations between the eight GATB aptitudes was subjected to a Schmid–Leiman hierarchical factor analysis. (Aptitude *G* [General Intelligence] was omitted from the present analysis, as it is not an independently measured aptitude, being a composite of the Verbal, Numerical, and Spatial aptitude scores.) The coefficient of congruence between the hierarchical *g* and the first principal factor is +0.998. How closely does *g* of the GATB obtained by our factor analysis correspond to the intelligence measured by standard IQ tests? The GATB *Manual* gives the correlations, in large adult samples, between each of the aptitude scores and total IQ (or some equivalent score) on each of 12 well-known standard tests of IQ or general intelligence. Presumably, such tests are largely measures of Spearman's *g*. (The 12 tests are the ACE Psychological Examination, California Test of Mental Maturity, Cattell Culture-Fair Test of *g*, Raven's Colored Progressive Matrices, the Reasoning Test of the Differential Aptitude Test battery, Henmon-Nelson IQ, Lorge-Thorndike IQ, Otis IQ, Beta, School and College Aptitude Test, Wechsler Adult Intelligence Scale, and Wonderlic Personnel Test.) The 12 correlations between each of the 8 GATB aptitudes and the 12 IQ tests were averaged (via Fisher's Z transformation). (Analysis of variance shows the profile reliability of this 8-point profile of averaged correlations is 0.96.) The correlation between this profile and the profile of Schmid–Leiman hierarchical *g* loadings is +0.85 (corrected for attenuation, +0.89). This means that the *g* factor of the GATB is highly similar to the general ability factor reflected in the total scores of standard IQ tests.

It is worth noting that although Aptitude *K* (motor coordination) has a *g* loading of .51 (and a mean correlation of .31 with 12 IQ tests), it shows nearly zero difference between the black and white means. This could happen only if black subjects, on average, were superior to white subjects on the non-*g* factor(s) (or specificity) measured by the motor coordination test.

### *Kaufman and Kaufman (1983)*

**SAMPLES.** The national standardization sample of the Kaufman Assessment Battery for Children (K-ABC) was used. The sample of children, ages 2 years 6 months through 12 years 5 months was selected in 1981 by a stratified random sampling procedure so as to be demographically representative of the U.S. population based on the 1980 U.S. Census results. Characteristics of the sample are described in detail in the K-ABC *Interpretive Manual* (pp. 62–71).

**ANALYSIS.** The present analysis is based only on the K-ABC standardization sample for school-age children, because the K-ABC includes a larger number (13) of subtests in this age range than for the preschool sample (10), and the sample size of the school-age sample is three times as large as that of the preschool sample.

The correlation matrix for the 13 subtests of the K-ABC in the entire ($N$ = 1500) school-age (5 through 12 years) standardization sample (combined population groups) is given in Table 4.11 (p. 92) of the *Interpretive Manual*. (Reliabilities of the subtests are given in Tables 4.1 through 4.4 [pp. 82–85]. The internal consistency [split-half] reliabilities are those listed in the present Table 5.) The correlation matrix was subjected to principal factor analysis, which provided the *g* for all the analyses used in Figures 1 and 2. However, in order to determine the similarity of the *g* factor in the black and white samples (for which correlations are not reported separately in the *Interpretive Manual*), the correlations and principal factors in the school-age group, separately for black and white samples, were provided by Dr. Cecil Reynolds (personal communication, July 1983), who is conducting detailed statistical analyses of the K-ABC standardization data. These *g* factor loadings for black and white samples are given in Table 5 and summarized in Table 3. The coefficient of congruence between the *g* extracted from the correlations based on the combined samples and the *g* extracted from the black and white samples separately are +0.999 and +0.997, respectively.

When Spearman's hypothesis is tested on just the eight mental processing subtests (i.e., excluding the achievement subtests), the correlation between the profile of standardized black–white differences on the eight mental processing subtests (i.e., the first eight K-ABC tests listed in Table 4) and their *g* loadings is +0.69, which shows that Spearman's hypothesis is borne out in the K-ABC regardless of whether the achievement battery (5 subtests) is included; inclusion of the achievement subtests in fact lowers the correlation between *g* loadings and the black–white differences to +0.58. A detailed critical review of the K-ABC with respect to the black–white difference has appeared elsewhere (Jensen 1984b).

The K-ABC *Interpretive Manual* gives the correlations of all 13 subtests with the WISC-R Full Scale IQ and with the Stanford–Binet IQ (p. 116). Since the WISC-R and Stanford–Binet IQ are commonly regarded as fairly good estimates of Spearman's *g*, it is worth noting the degree of relationship between the profile of correlations of each of the K-ABC subtests with the WISC-R IQ and Stanford–Binet (S-B) IQ and the profile of *g* loadings of the subtests in the present factor analysis for the total school-age sample. The correlations between the profiles are as follows: WISC-R × S-B = +0.85; WISC-R × *g* = +0.79; S-B × *g* = +0.83. In brief, the *g* of the K-ABC is highly similar to the *g* of the WISC-R and Stanford–Binet, even though the item contents of these three batteries are all quite diverse.

### *Veroff, McClelland, and Marquis (1971)*

**SAMPLES.** Black ($N$ = 186) and white ($N$ = 179) adults between 18 and 49 years of age were randomly selected from a probability sample of 1,027 households within the city of Detroit, sampled so as to yield a cross section of each population. Six ability measures were administered to approximately half of each sample by either a black or a white interviewer.

Principal factors were extracted from the intercorrelations of the six tests in this study. Although Veroff et al. do not present means and standard deviations for each population, they report an analysis of variance on each test showing the mean squares between and within populations and the $F$ ratio for the population main effect. The mean black–white differences (in $\sigma$ units) can be calculated from these statistics. Of course, the rank order of the population $F$ ratios is exactly the same as the rank order of the standardized mean black–white differences on the six tests.

### *Hennessy and Merrifield (1976)*

**SAMPLES.** The subjects were high school seniors planning to enter an open-admissions community college in the City Uni-

versity of New York. The white sample used in the present study does not include persons of Jewish or Hispanic background.

The correlation matrices for the black and white samples, given in Tables 1 and 2 of Hennessy and Merrifield (1976), were subjected to a principal factor analysis. The means and standard deviations of the black and white samples on the 10 tests are found in a doctoral dissertation by Hennessy (1974). Reliabilities are reported for only 7 of the 10 tests.

Because Hennessy and Merrifield were primarily concerned with various ethnic population differences in the factor structure of abilities, they partialed socioeconomic status (SES) out of the correlations among all of the ability measures. The SES index was based on family income, the occupation of the main wage earner, and the educational level of both parents. (When SES was included in the factor analysis, it showed no loadings above 0.20 on any of the three ability factors that emerged. This small effect of SES, however, was statistically removed from the factor analyses used in the present study.)

## NOTES

**1.** Throughout this paper, the black–white difference is always expressed as the white mean *minus* the black mean, divided by the square root of the *N*-weighted average variance within the two groups.

**2.** The coefficient of congruence, $r_c$, is an index of factor similarity on a scale of 0 to ±1. Unlike the Pearson $r$, which, being based on standardized variates, reflects only the degree of similarity between the profiles (of factor loadings) per se, the congruence coefficient also reflects differences in the absolute values of the factor loadings. A value of $r_c$ above +0.90 is the usual criterion for concluding identity of factors, although some experts set a more stringent criterion at +0.95. The congruence coefficient is computed as follows:

$$r_c = \frac{\Sigma ab}{\sqrt{\Sigma a^2 \Sigma b^2}}$$

where $a$ and $b$ are the homologous factor loadings obtained on a given factor in groups $A$ and $B$.

**3.** I am indebted to Professor John Schmid for performing all three of the hierarchical factor analyses used in this paper.

**4.** The correlation between $g$ loadings and socioeconomic status (SES) within populations has no direct relevance to Spearman's hypothesis, which concerns only the difference between black and white populations. To the extent that $g$ is a strong selective factor in occupational status attainment, however, one should predict a positive correlation between various tests' $g$ loadings and the magnitudes of the average SES differences on the tests within either the black population or the white population. In a study by Jensen and Reynolds (1982), the rank-order correlation (Spearman's rho) between 13 WISC-R subtest $g$ loadings and SES differences (i.e., Pearson correlation between SES classified on a five-point scale and subtest score) within the total white standardization sample was found to be +0.73 ($p < .01$); the corresponding correlation within the total black sample was +0.57 ($p < .05$). The fact that tests' $g$ loadings are correlated with SES differences as well as with black–white differences has no direct bearing on the validity of Spearman's hypothesis, however. The Spearman hypothesis pertains only to the psychometric nature of the black–white difference on various tests and in no way addresses the cause of such differences.

**5.** One of the referees of this article has suggested that one "dissenting study" (Humphreys, Fleishman & Lin 1977) is omitted from consideration in the present analyses. The study, as presented in the article by Humphreys et al., however, is unsuitable as a test of Spearman's hypothesis for several reasons. In the first place, the Humphreys study does not present the basic elements needed for a direct test of the hypothesis, namely, means and *SD*s of representative black and white samples and $g$ factor loadings of the various tests, in this case the large battery of tests used in Project TALENT. The Humphreys data consist entirely of school means and have not been analyzed at the level of individual differences; $g$ factor loadings of the tests are not reported. Moreover, the nature of the Humphreys data would not permit extraction of factors comparable to those considered in the present analyses, all of which are based on factor analyses of individuals. An even more serious objection is that, in the Humphreys study, comparisons of the profiles of test means are made between a black sample and either a low-socioeconomic-status (SES) white sample or a high-SES white sample; no comparisons are made between the black sample and a representative white sample including all SES groups. This violates one of the methodological desiderata listed early in the present paper, namely, that the black and white samples should not be selected on the basis of any variables that are themselves highly related to $g$. SES is notably correlated with $g$. The suitability of some of the tests in the Project TALENT battery may also be questioned as a vehicle for testing Spearman's hypothesis. Many of these tests are very short, relatively unreliable, and 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 that there is a "floor effect" for the black sample. That is, the items are too difficult to allow measurement of the full range of ability in the black sample, a phenomenon that has the effect of spuriously diminishing the observed difference between the black and white means. Readers are urged to read the article by Humphreys et al. (1977) in order to judge for themselves the claim that it contradicts Spearman's hypothesis. In the present writer's judgment, these data, at least in the form in which they are presented by Humphreys et al., support no worthy inference vis-à-vis Spearman's hypothesis.

Another referee has suggested that some studies by Sandra Scarr might contradict Spearman's hypothesis. A search through all of Scarr's published empirical studies in which there are comparisons of black and white groups has turned up one study with some direct relevance to the hypothesis. This study was not included in the present analyses because one of the cutoff decisions for including studies was that they have used at least six different tests, so as to permit a reasonable factor analysis and range of factor loadings. Scarr's (1981b, pp. 261–315) study involves only five tests: Raven Matrices, Columbia Mental Maturity Scale, Peabody Picture Vocabulary, Benton Visual Retention Test (conceptual memory for designs), and a paired-associates rote learning task. Scarr (Table 8) presents black and white means and *SD*s on each of the tests, based on good-sized *N*s (183 to 447), and $g$ loadings (i.e., first principal component, Figure 11.4.1). Hence we can treat Scarr's data in exactly the same fashion as the other data sets were treated, that is, correlating the mean black–white differences (in $\sigma$ units) on the five tests with the tests' $g$ loadings. The results are shown below, with the $g$ loadings derived from the white ($g_W$) and black ($g_B$) samples reported separately.

| *Test* | *W-B Diff.* | $g_W$ | $g_B$ |
|---|---|---|---|
| Raven | .91σ | .80 | .82 |
| Columbia | .63σ | .77 | .74 |
| Peabody | 1.15σ | .70 | .76 |
| Benton | .65σ | .74 | .77 |
| Paired-associates | .36σ | .60 | .50 |

The correlation of the W-B Diff. with $g_W$ is $r = +0.46$ ($\rho = +0.30$), and of the W-B Diff. with $g_B$ is $r = +0.73$ ($\rho = +0.70$). The average $r$ is +0.61. Thus, Scarr's study is quite in line with the correlations obtained in studies employing a larger number of tests, the mean correlation for which is +0.60. An obvious

limitation of Scarr's brief test battery is that all of the tests except paired-associates have such high and similar *g* loadings as to greatly restrict the variability upon which the test of Spearman's hypothesis depends.

# Open Peer Commentary

*Commentaries submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.*

## Jensen, Spearman's *g*, and Ghazali's dates: A commentary on interracial peace

Panos D. Bardis

*International Social Science Review, University of Toledo, Toledo, Ohio 43606*

*The soul, besides other things, contains intelligence, and the head, besides other things, contains sight and hearing; and the intelligence mingling with these noblest of the senses, and becoming one with them, may be truly called the salvation of all things.*

**Plato,** *Laws*

*Intelligence is one and continuous, like thought.*

**Aristotle,** *De Anima*

***1. Introduction.*** The author has something to say. The commentator has to say something. Accordingly, the latter's task is more difficult, especially when space is limited and the author is Arthur Jensen, who now claims that psychometric, chronometric, and other tests support Spearman's *g* hypothesis.

Since Jensen's recent work on Spearman's *g* hypothesis is original, since his methodology is basically sound, and since his emphasis is on truly scientific research, the objective commentator cannot dismiss him. But the critic must not ignore humanitarian issues either, since the practical implications of Jensen's conclusions concerning blacks are devastating. To me, humanitarianism is exceedingly important. But more important are truths such as these: "Attitudes are not innate" "The solar system is heliocentric," and "If $p$ and $a$ are positive integers, $p$ is a prime, and $a$ is prime to $p$, then $a^{P-1}$ divided by $p$ leaves a remainder of 1." For a humanitarianism that disregards such truths becomes nothing but a dangerous sentimentality. Therefore, if Jensen had been more convincing, I would slight humanitarianism in this sphere and pursue its goals in some other fashion. But more fundamental and philosophical arguments generate certain doubts regarding Jensen's admittedly impressive work, thus "equalizing" it with humanitarianism in a way that partly recalls the two dates of which the Arab philosopher Ghazali (1058–1111) wrote. These extremely similar oblong fruits, placed in front of a hungry man who was equally attracted to both of them, made it difficult for the man to select one of them, since he was unable to take them both.

Below I will discuss selected issues that lead to such indecision.

***2. Spearman's genius.*** We cannot dismiss Spearman. His concept of "general intelligence," his "neogenetic laws," the Spearman rank-order correlation coefficient, the Spearman–Brown prophecy formula, and his factor analysis will not allow us to do so.

***3. g and Plato's third-man argument.*** But what are the problems involved in *g*?

Well, these problems concern the nature of *g* itself. First of all, we should state that *g* is the "eduction" of a relation between two entities, which Spearman himself called "fundaments." Its mathematical expression is as follows:

$$z_{ij} = a_{jg}z_{ig} + a_{js}z_{is}$$

Here *g* is the universal factor and *s* a component unique and specific to test *j*.

Some of the main criticisms may now be outlined as follows:

1. Perhaps Spearman's initial methodological and ontological reductionism is questionable.
2. What we really need to do is to discover the essential nature of *g* in terms that are independent of factor analysis.
3. Even one genuine zero correlation between pairs of intellectual tests would prove the nonexistence of a universal factor such as *g*.
4. I have always suspected that it was philosophy, Spearman's first love, that led him to the discovery of a universal *g*. But the history of universals indicates that this philosophical concept is too problematical. Consider, for instance, Plato's *eidos* and *Idea*, Aristotle's "*ta katholou*," the medieval *universalia*, Locke's "abstract ideas," Hume's "resemblances," and Wittgenstein's "family resemblances."

Plato, the father of universals, deserves additional attention in this context. In his dialogues, while discussing *eidos* and *Idea*, Plato looked for a general entity and a general word name for it. This he considered necessary both ontologicaly and epistemologically. Needless to add, he did not wait for Aristotle to criticize his theory. Toward the end of his life, Plato wrote his *Parmenides*, in which he vacillated between the belief that his theory of Forms was perfect and the problems with it that he himself had stated and been unable to solve. For instance, according to his third-man, or infinite-regress, argument, since all particulars are merely imperfect copies of a perfect Form, and since a Form is one over many particulars, the Form shares a feature with its particulars. But this feature necessitates the existence of another Form and so on ad infinitum. This was Plato's Waterloo!

***4. The nature of intelligence.*** Nowadays interracial peace and harmony are also influenced by researchers' conception of intelligence in general. Of course, beginning with the Binet–Simon instrument of 1905, progress in this area has been spectacular. We cannot disregard Thurstone's multivariate approach, Wechsler's "nonintellective factors," Guilford's "structure-of-intellect," and so on. But relative chaos is still prevalent.

I am convinced, therefore, that greater progress will be achieved when psychologists begin to imitate physicists. Unfortunately, so far, too many psychologists have, instead, thought of factors as imaginary abstractions and of multiple factor analysis as synonymous with faculty psychology. They obviously forget that discovering a functional unity by means of correlation has nothing to do with inventing a faculty and attaching a label to it.

***5. Intelligence tests, the bare bear, and the great "circulator."*** One remains equally skeptical and ambivalent when it comes to the instruments that measure intelligence.

Not only has sampling often been unrepresentative; conventional tests have also stressed "convergent thinking," thus neglecting creativity. Recent instruments, which emphasize "divergent thinking," are more satisfactory, but they, too, have their limitations. When such raw data constitute the foundation for advanced statistical tests, how valid and reliable can the conclusions be? As a British statesman observed, Her Majesty's statistics are as good as the data collected by the least constable at the local level.

As for black–white differences, two major hypotheses have thus far been formulated in order to explain this gap: that the

tests are linguistically biased and that they are culturally biased. Arthur Jensen asserts that all relevant empirical studies have rejected both hypotheses. And, at the present time at least, it seems difficult to refute his assertion. But when one scrutinizes "culture-free" tests, "culture-fair" instruments, and so on, one cannot agree with Jensen either.

Consider this test item: "A. Bare, B. Bear, C. Hare, D. _____. "Think of the A–B and C–D relationships. D should be assigned one of the following words: "Hair," "Hear," "Hairless," and "Tolerate." Suppose upper-class blacks said "Hair," upper-class whites "Hear," lower-class blacks "Hairless," and lower-class whites "Tolerate." Which answer would Spearman and Jensen say is correct? I would say, all of them! Just look at the four words.

In brief, who is to judge? How? Why?

William of Occam (1285–1349) stated the law of parsimony: "*Entia non sunt multiplicanda praeter necessitatem.*" In a study of various tests, I formulated the following principle: "*Instrumenta scientiae non sunt involuta praeter necessitatem*" (Bardis 1969). Valid, reliable, parsimonious, accurate, rigorous tests will certainly promote our scientific knowledge.

Finally, we must not condemn currently unorthodox attitudes and abilities, since they may be indicative of and conducive to genuine creativity. Most biologists still do not realize that William Harvey's (1578–1657) spectacular achievement was primarily due to the *mystical* atmosphere at the University of Padua, where he received his medical degree, not to unadulterated empiricism, experimentalism, and inductive reasoning – these attitudes prevailed only after the Padua period, by which time Harvey had returned to London. On the contrary, during Harvey's studies Padua was dominated by the Heraclitean–Platonic theory of *cyclical* universal evolution and by the belief that the *microcosm* of man is a replica of the *macrocosm* of the universe. The *circulation* of the blood thus appeared to be a logical conclusion. But Harvey was disparagingly nicknamed "circulator" (mountebank).

Indeed, who is to judge? How? Why?

**6. *Nature versus nurture.*** The most controversial issue in this area is that of genetic and cultural causation.

Arthur Jensen himself stated in 1977 that the IQ of black children increases with age in California but decreases in Georgia (Jensen 1977b). In New York, black and Puerto Rican children perform better on tests if they have resided longer in that city. In 1980, Hunt reported that life in an orphanage tends to result in slow intellectual development. In high school, the IQ of students who take science and mathematics increases, while that of students who take domestic "science" and dramatics decreases. Environment, then, does seem to be influential.

But so is heredity. Intelligence correlations, for example, are about .90 for identical twins and .50 for siblings. The Wechsler Adult Intelligence Scale favors the male to a slight extent – the reasons are obviously cultural. And so on.

In brief, although Vernon (1979) attributes 60% of intelligence to heredity, 30% to environment, and 10% to their interaction, even the most impressive findings remain inconclusive. Of course, this is not surprising, since environmental differences have not been quantified adequately as yet, and since both hereditary and environmental effects have been treated primarily summatively, not interactively – which, admittedly, is exceedingly difficult.

**7. *Conclusion.*** Peace and war are not exclusively international phenomena. They can be internal (psychological) or external (social). They can involve individuals or groups. And they can be of any degree. Accordingly, we cannot ignore the implications of Jensen's research. He himself states that, statistically speaking, blacks will have a greater handicap in those educational, occupational, and military spheres that are highly correlated with *g*. So is Jensen's work a new Pandora's box? And must we suffer the same fate as Prometheus or Epimetheus? Are Jensen's unquestionably admirable investigations a new magic broom that will create a devastating cataclysm from which neither Goethe's noble lyrics nor Dukas's beautiful melodies can rescue us? Who knows?

In 1969, Jensen began to stress genetic, environmental, and cultural factors in order to understand individual and population differences. What is so monstrous about that? Perhaps nothing. However, it is monstrous to attempt to silence him, as many have often done. So we have to choose between academic freedom and research implications.

And now, back to Ghazali's dates. After this detailed analysis, I still feel like the Arab philosopher's proverbial man, hoping that Jensen's own future research will soon prove that there are no black–white differences in *g*. Of course, fanatics on either side will pejoratively whisper something about Buridan's ass. My first answer to them would be that, like the French scholastic philosopher's enemies, they are at least careless. Jean Buridan (1295–1356) never mentioned such an animal. Inspired by Aristotle, he only wrote, in his *Expositio Textus*, about a perplexed and puzzled pooch between two equal portions of food. Then, I would refer them to Aristotle, who, in his *De Caelo*, describes "the man who is fiercely and equally hungry and thirsty, and stands at an equal distance from food and drink; and for whom it is therefore necessary to remain motionless" (295b).

## Reliability and *g*

Jonathan Baron

*Psychology Department, University of Pennsylvania, Philadelphia, Pa. 19104*

Reliability can affect both a test's *g* loading and its power to discriminate groups (in terms of standard score units). Jensen disputes the hypothesis that his findings could result from differences in test reliability, but several points argue in its favor:

1. The relevant reliability measure is neither split-half reliability (reluctantly used in most cases here) nor stability (test–retest reliability on the same items) but rather test–retest reliability with parallel forms of the test. Both lack of generality over items of the same type and lack of stability over time could reduce a test's *g* loading or its power to distinguish groups. Because this type of reliability (were it known) is likely to be lower than split-half reliability (or stability), correction of *g* loading (or difference score *D*) by disattenuation (Jensen's Table 3) is likely to be an undercorrection.

2. Correction by partialing is also likely to be an undercorrection, because the partialed variable (reliability) is *inaccurately* measured (by split-half reliability).

3. In studies J, R, and K, reliabilities (*r*) correlate as highly with *g* loadings (Gw and Gb) as these correlate with each other, raising the question of whether reliability and *g* loading can be distinguished at all. (The relevant correlations are .85 for *r* and Gw, .82 for *r* and Gb, .89 for Gw and Gb, for studies J and R; .94, .82, .92, respectively, for K.) The role of reliability in other studies of *g* remains an open question.

4. In the remaining studies with at least seven reliabilities reported, *r* correlates about as highly with *D* (white–black difference) as do Gw and Gb, and the correlation of *r* and *D* is positive when the *g*-loadings are partialed, in all but one study–see Table 1. (The reliabilities of tests 6 and 11 are actually stability coefficients, unlike the other measures reported. When these tests are omitted, the results fall more closely into line with the hypothesis that reliability, not *g* loading, is the main determinant of *D*; see the rows marked with * in Table 1.)

5. Even if reliability as we know it cannot explain the results, there is another type of reliability to consider, the extent to which the items on a test can predict performance on items of the same *general* type. Thus, the reliability of the digit span

Table 1 (Baron). *Correlations relevant to the comparison of reliability (r) and g loadings (Gw and Gb) as predictors of black–white difference (D)*

| Correlation: | *D*,*r* | *D*,Gw | *D*,Gb | *D*,*r*/Gw | *D*,*r*/Gb | *D*,Gw/*r* | *D*,Gb/*r* |
|---|---|---|---|---|---|---|---|
| Sandoval (12) | .43 | .36 | .50 | .27 | .15 | .08 | .32 |
| Sandoval (10*) | .41 | .09 | −.01 | .47 | .61 | −.27 | −.50 |
| Mercer (12) | .52 | .66 | .66 | .12 | .13 | .49 | .50 |
| Mercer (10*) | .53 | .37 | .32 | .41 | .46 | .00 | −.10 |
| Nichols (7) | .78 | .73 | .74 | .75 | .67 | .69 | .59 |
| D. of D. (10) | .39 | .40 | .29 | .29 | .40 | .29 | .29 |
| D. of Labor (8) | .53 | .71 | — | .29 | — | .62 | — |
| Hennessy (7) | −.45 | .14 | −.06 | −.46 | −.47 | .20 | −.03 |

*Note:* Asterisks indicate that tests 6 and 11 have been omitted. Partialled variables appear to the right of a slash.

should be measured not by using a parallel *digit* span test, but by using other kinds of span tests, such as letter span, word span, and so on. Possibly, the more *g*-loaded tests are simply those consisting of items more broadly sampled from their general class (assuming that this could be defined). A broader test would be more sensitive to a group difference within the entire class it measures, for it is less affected by idiosyncratic individual differences in specific tasks.

Jensen suggests that *g* is a single source of variance in test items and must be explained primarily in terms of physiology rather than learning history. This hypothesis is supported by the correlations with reaction-time measures and evoked potentials. However, the evoked potential might be simply an index of attention, or some other single factor other than *g* that affects many tests. Reaction time may also be sensitive to such a factor, and it may also be influenced by preparation, motivation, learning, vigilance, and fatigue even in the simplest tasks. (Jensen, 1982b, reports that in his reaction-time tasks, the simpler conditions are always run first, so that slope differences may result from practice, vigilance, or fatigue effects within the session, for example.) The results of Jensen's Figure 10 (and similar results) could be explained in terms of the influence of such factors for the high-latency tasks, and the larger proportional contribution of perceptual and motor processes to the low-latency tasks. The results of Figures 11 and 12 (and similar results) could be due to a scaling problem: The longer the latency, the more room there is for any variable to affect it (Baron & Treiman 1980).

Jensen also suggests that *g*-loaded tasks require more steps and more parallel processing. It is not obvious that this sort of account will work. Memory-span tasks, for example, can be set so that they require considerable parallel processing (e.g., holding some digits in one store while rehearsing other digits in another). Forward and backward span tasks need not differ in the number of operations or in the extent to which parallel processing is involved; backward span does require reversing the digits, but fewer digits are involved.

Supposing all my criticisms so far to be wrong, let me suggest another hypothesis about *g*. Within tasks of the narrow type used in IQ tests (see Baron, in press, Chap. 1), the less *g*-loaded ones have little in common, but the more *g*-loaded ones have *two* attributes in common, which are correlated across the tasks. First, they are more likely to lead to errors in which the subject does a different (more natural?) task with the same stimuli, such as providing an association rather than a definition or analogy, or tending to recall forward rather than backward. The attention and self-control required to avoid such errors may have physiological determinants, and these would account for the correlations with physiological measures and for evidence of heritability. Second, *g*-loaded tasks are more sensitive to what I have called cognitive style (Baron, in press; this is similar to what Jensen calls metaprocesses). That is, *g*-loaded tasks require thinking, considering alternative possibilities, and gathering and using evidence. This attribute might be sensitive to cultural differences in the encouragement of caution and self-criticism as opposed to quickness and bravado. Even the backward digit span might be more sensitive to such stylistic factors than the forward span, for it might be worthwhile in this task to check to see that one has learned the string well in the forward direction before trying to reverse it. Such a confounding of test attributes would be consistent with the existence of both physiological and cultural effects on *g*.

## Looking for Mr. Good-*g*: General intelligence and processing speed

John G. Borkowski and Scott E. Maxwell

*Department of Psychology, University of Notre Dame, Notre Dame, Ind. 46556*

Jensen has marshalled evidence in support of the argument that the major source of black–white differences in IQ is Spearman's *g*. Two issues follow closely upon his initial observation: How do we proceed to endow *g* with meaning? What theoretical and methodological pitfalls confront those who persist in the search for its elusive nature?

The history of the psychometric approach to intelligence conveys a harsh fact. Spearman's *g* is a creature of statistics, possessing no theoretical import. It fails to yield explanatory insights. It provides little or no direction for future research or for theory construction. No wonder that Jensen — following his observation about the relation of *g* to black–white differences in IQ — would continue searching for the nature of *g*. The flow of research events in this tradition proceeds in an orderly fashion: from the construction of a battery of tests on some logical grounds, to the calculation of *g*, to the identification of new tests that correlate with an index of *g*. In the present instance, speed or rate of elementary information processing is identified by Jensen as a major correlate of *g* and a somewhat smaller but important correlate of black–white IQ differences.

Jensen's indirect approach to theory development poses several potential problems. First of all, the relationship of rate of processing to *g* takes on clear, unambiguous meaning only when it is contrasted with other potential correlates such as processing skills, metacognitive states, and domain-specific knowledge. Since each of these factors has been postulated as important to intellectual performance (Borkowski 1985), they stand as viable candidates against which Jensen's notions about "speediness"

can be falsified. If rate of processing clears the hurdle when tested against other potential explanatory factors, it could take on greater exclusivity as the important underlying source of *g*.

In the reliance on *g* to validate new theoretical ideas, the multidimensional, developmental character of components is obscured, if not lost altogether. Sternberg (1984a), Horn (in press), Butterfield and Ferretti (in press) and others have, in recent years, brought to our attention critical issues about how components of intelligence interact and how the nature of these interactions changes with age. Jensen recognizes the first of these points in suggesting that metaprocesses might control the deployment of elementary processes, contributing directly to variations in *g*. This view is in line with our recent work suggesting that metacognitive and process differences are associated with the "typical" black–white IQ differences (Borkowski & Krause 1983).

The second issue, however, is even more critical for theory construction: What is the pattern of development for the various components of intelligence (or cognition) and how will we proceed to study these developmental patterns? On the theoretical side, hypotheses are required about timing and sequencing as well as about why unique interactions of components change with various stages of mental development. For instance, Borkowski and Peck (in press) have speculated that elementary information processing guides the outcome of early performance in gifted children and also alerts parents to the fact of "giftedness." Early forms of parental stimulation in turn accelerate the emergence of metacognitive knowledge that is the setting condition for reflective, strategy-based problem solving in middle childhood. The static concept of *g* actually seems to hinder this type of theorizing about the development of intellectual components.

*On the methodological side, recent developments in structural* equation modeling would allow Jensen to test explicitly the "strong" form of Spearman's hypothesis (cf. Rock, Werts & Flaugher 1978). The point here is that Jensen's approach cannot address the question of whether black and white populations differ only on *g*. In fact Jensen admits that an inspection of mean differences in several of the data sets contradicts the strong form. But what about the "weak" form of Spearman's hypothesis that holds that "the black–white difference in various mental tests is *predominantly* [emphasis ours] a difference in *g*"? Although Jensen demonstrates that tests which load most highly on *g* tend to show the biggest black–white difference, nowhere is it shown that this difference is *predominantly* a difference in *g*; the "weak" hypothesis remains untested. The factor analysis of preexisting test batteries obviously provides no hints as to why alternative sets of interrelationships might arise in the developing organism. We believe that structural equation modeling will prove more useful in the investigation of complex, longitudinal, causal relationships among multiple cognitive constructs, permitting tests of the "weak" form of Spearman's hypothesis.

There is reason to question the "purity" (or construct validity) of Jensen's primary measure, rate of information processing. For instance, Jensen (1980a) presented data from Noble (1969) on changes in reaction time (RT) for black and white children across a sequence of trials. Although no RT differences were observed initially, RTs improved more rapidly for white children than for black. Two points are noteworthy here: (a) The fact that RT increased with practice suggests the presence of skill components that develop and, presumably, interact with stable elementary processing components in influencing performance. (b) Although Jensen (1980a) concluded from the Noble data that differential motivational factors are absent on RT tasks, it seems more plausible to suggest that variables correlated with socioeconomic or black–white differences, such as perseveration, attributional beliefs, and locus of control probably influence RT performance, especially on later trials that demand attention and vigilance in the face of boredom and distraction. In a similar vein, Carlson and C. M. Jensen's (1982) investigation of the relationships among reaction time, movement time, and Raven scores led them to conclude that "some factor or group of factors other than information-processing capacity or speed are involved in the relationships observed. One of the factors may be motivation, or a tendency to want to perform well" (p. 272). In short, there is reason to question RT slopes and RT variability as "pure" measures of rate of elementary processing. Personality-motivational factors and acquired skills probably influence RT performance.

There are final reasons, with educational relevance, that need to be considered in arguing against the use of *g* as a research framework: Intellectual components have unique origins and differential degrees of modifiability. Horn (in press) has argued persuasively that distinct intellectual factors (e.g., $G_f$ and $G_c$) have independent developmental trajectories and different degrees of heritability. Hence, it makes little sense to speak of the heritability of Spearman's *g* or to struggle with an analysis of its determinants.

Multidimensional perspectives on intelligence not only allow for theoretical diversity in understanding how components emerge, grow, and decline but also invite training studies designed to assess their degree of modifiability. From this framework, we can speculate about how much particular skills, metacomponents, or pieces of knowledge influence learning or problem solving in both applied and laboratory settings. Finally, we can determine whether, and how much, the remediation of intellectual (or cognitive) deficits affects academic performance. The multicomponent approach to intelligence, perhaps couched in the mold of dynamic assessment (Day, French, & Hall 1985), holds promise for simultaneously testing ability and influencing academic achievement (cf. Palinscar & Brown 1984).

*We return, then, to the title of this commentary: Should* research on IQ continue to chase after Mr. Good-*g*? Our view should be clear: There are attractive alternatives to Mr. Good-*g* who might be more suitable companions in the search for the nature of intelligence.

ACKNOWLEDGMENT

The writing of this commentary was supported, in part, by NIH grant HD-17648.

## Jensen's compromise with componentialism

Christopher Brand

*Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, Scotland, U.K.*

Jensen's article is both scholarly and powerful: with all the skill and assiduity of the world's most impressive psychometrician he mounts an argument that should subdue objection and compel assent. However, I think I know my experimentalist colleagues well enough to predict that they will not be overawed by Jensen's heroic effort to leap the barrier between Cronbach's (1957) two psychologies while saddled with a relatively biological interpretation of black–white (B–W) differences in IQ.

Jensen's use of hierarchical factor analysis whenever possible is a notable advance and may serve to remind psychometricians of what they have been missing. Factorists, in their determination to elicit what is laughably called "simple structure" from intrinsically messy psychometric tests, have been unduly prone to rotating *g* variance away into the alleged specifics of which they have held intelligence to be composed: so it is high time to use a method of factoring that allows *g* its rightful place. Such a move is particularly necessary after many years of scientific failure to obtain adequate, differentiated accounts of Thurstone-type components of intelligence (e.g., Scarr & Carter-Saltzman 1982).

Apart from highlighting the basic nature of the B–W difference, it is of particular interest that, once *g* is partialled out, whites have only a relatively small advantage in performance (or spatial?) ability; and that by the same procedure blacks emerge with an advantage in memory, as Jensen has long envisaged. One only wishes that students of individual differences would take to using hierarchical factor extraction methods more widely when dealing with human personality and attitudes: we might then see more clearly what were the real, discrete, non-*g* sources of variance in dimensions that have arguably suffered from too much confusion with *g* and its educational sequelae – the study of "authoritarianism" springs to mind (Brand 1984; 1985a).

At the same time, I fear that experimentalists – always prone to environmentalism by the nature of their manipulative, all-things-are-possible, utopian trade – will find good sport in teasing out the implications of Jensen's compromise with componentialism. It may well be that psychometric *g* – established hierarchically or by whatever (to experimentalists) arcane procedure – encapsulates the very quintessence of the B–W difference. But what does that matter if *g* itself is held by Jensen to be dissoluble into a multiplicity of abilities on elementary cognitive tests (ECTs)? After all, must not any serious, modern environmentalist explanation of the B–W difference involve a number of environmental differences having a host of small effects on each of the legion of black boxes of which the modern experimental cognitivist holds the mind to be composed? If, at last, psychometric intelligence turns out to call to some degree on many of these black boxes – on short-term memory, long-term memory, immediate memory, working memory, and so forth – then is the road not still open to environmentalist explanation of the B–W difference?

Moreover, Jensen's admission here that each of his ECTs on its own has but a modest correlation with IQ will particularly strengthen the hand of the "wetter" (as is said in Britain) cognitivists. Such theorists will readily attribute Jensen's low correlations not to "basic processes" being the *fons et origo* of *g* but to the influence of developed intelligence on ECT performance resulting from the use of slightly superior "strategies" by brighter, test-wise subjects in the laboratory. All in all, Jensen's psychological componentialism must tend – though I only say "tend" – to undermine his psychometric unitarianism as to the explanation of the B–W difference.

From this side of the Atlantic, it is clear that the humble endeavours of – shall I say – Commonwealth psychologists to advance a psychologically unitarian account of *g* have so far met with little approval from our cousins across the water. Predictable though it may be, I have to say that measures of "inspection time" (IT) for extraelementary displays continue to show strong correlations (of around −.60) with measures of IQ and mental age (e.g., Brand 1985b). The most recent Scottish study, by Donald Sharp (1984), on adolescent children of mean IQ 103 (with a standard deviation of 12 points – a somewhat restricted range) gave an uncorrected correlation of −.54 ($p<.01$) between tachistoscopic IT and Raven's matrices. Such work, I surmise, suggests the possibility of identifying a really substantial proportion of variance in fluid *g* as "mental intake speed."

Still more seriously – and expertly, in terms of the methods involved – the Hendricksons' (1982) work with their string-length measure of average evoked potential (AEP) sits quite unchallenged in the literature, apparently defying the critical faculties of leading American environmentalists who are certainly familiar with it. Although the Hendricksons' biochemical theory (of how intellectual differences arise from differences in synaptic transmission processes) may raise some eyebrows, their psychophysiological effect appears robust: the correlation of .80 between string-length AEP and IQ has been successfully replicated by Ian Fraser (1984) in Edinburgh on a student sample ranging down only to IQ 100. While Jensen mentions the work of Nettelbeck (in Adelaide) and the Hendricksons, his subsequent preference for componentialism suggests that he has – strangely for such a nonconformist martyr to the social sciences – decided to settle for conventional wisdom. As for Jensen's invocation of "working memory" as an explanatory variable, I can only hope he will look into Dempster's (1981; 1985) work with its implication that speed-of-intake differences account for quite a lot of the variance that is seen in the laboratory of the experimental student of memory's black boxes.

Of course, there are "strategies" that are sometimes used successfully by subjects in IT as well as RT studies – particularly when visual stimuli are presented on the TV screen or by means of light-emitting diodes; so, just conceivably, there may even be strategies that are used in AEP studies by subjects who are told, in a darkened room, "Close your eyes, relax, and think of nothing in particular." But it has yet to be shown that strategies themselves account for the striking correlations with IQ that these IT and AEP procedures have generated: rather, the use of strategies seems only to weaken such correlations. Maybe Jensen knows something that I don't know about these kinds of procedure. If so, I hope he'll be frank in his reply. Otherwise, I would still hold out the simplistic hypothesis that fluid *g* is not only psychometrically but also psychologically unitary, and that the developmental basis of differences in intelligence consists in differences in ability to readily apprehend even the most elementary features of the real world.

## Event-related potentials and the biology of human information processing

Enoch Callaway

*Langley Porter Psychiatric Institute, University of California, San Francisco Medical Center, San Francisco, Calif. 94143*

I am sorry that Jensen made an unnecessary excursion into psychophysiology to support his position. He makes his basic points quite adequately with factor analyses of conventional test scores. The notion of a general intelligence factor is well supported. The fact that intelligence in humans has a biological basis seems too obvious to merit serious discussion, although I can understand why Jensen might be inclined to belabor the point.

I share his interest in the biological mechanisms that underlie *g*, but the way he uses studies of brain electrical potentials to justify that interest could be misleading. This forces me to point out that short event-related potential (ERP) latencies do not necessarily go with intelligence and that brain electrical potentials are not necessarily more "biological" and less "psychological" than other sorts of human behavior. Only then can I get to the more interesting topic of how one might study the biological basis of *g*.

When Ertl and Schafer (1969) proposed the notion of "neural efficiency," it did not seem too unreasonable. Fast neurons might well be associated with fast behavior. However, now we know that *g* is associated with choice reaction times, and not with simple reaction time or tapping speed. In 1960, it was also not too naive to think that short-latency ERPs might likewise reflect fast neurons. We (Callaway 1975), among others, confirmed Ertl's findings of negative correlations between IQ and visual ERP latency.

The P3 is a positive component of the ERP. It occurs from about 300 to 600 msec after stimuli, particularly if they are relatively rare and paid attention too. It is relatively independent of stimulus modality and seems to reflect time taken in stimulus evaluation rather than in response selection (McCarthy & Donchin 1981; Duncan-Johnson 1981). P3 latency has been reported to correlate negatively with digit span (Polich, Howard & Starr 1983). Old demented subjects have longer P3 latencies than do nondemented older subjects. I suppose dementia could

be considered a specific disorder of *g*. However, there are so many other things that affect P3 latency that it is not of much clinical value in the diagnosis of dementia (Pfefferbaum, Wenegrat, Ford, Roth & Kopell 1984).

Jensen, however, fails to remark on examples of positive ERP latency/IQ correlations that would seem contrary to his position. Since the expected negative correlations do not necessarily support a biological interpretation of *g*, by the same token positive latency/IQ correlations do not suggest that *g* has no biological basis. Thus, even though Jensen's psychophysiological arguments seem irrelevant to the important substance of his paper, it would be misleading to leave the impression that negative ERP latency/IQ correlations always occur. For example, in spite of the well-established relationships between long-latency P3s and dementia, Ray Johnson (personal communication) has recently obtained positive correlations between P3 latency and IQ.

There are also other contrary examples. Hendrickson and Hendrickson (cited in Weiss, 1984) report negative correlations between auditory ERP latencies and IQ. E. Hendrickson was kind enough to send us some of her preliminary data in 1974, and we tried to replicate her findings. To our surprise, we found a positive correlation between auditory ERP latencies and IQ. We also noted that Ertl (1969) had reported a similar finding in adults, and that both Straumanis, Shagass & Overton (1973) and Hogan (1971) reported shorter latencies in retarded children than in controls. This is discussed at greater length in Callaway (1975). For our purposes here, it is enough to note that short ERP latencies do not necessarily indicate intelligence.

The existence of both positive and negative ERP latency/IQ correlations is not surprising considering how many factors influence ERP latencies. Among the many things that can affect P3 latency are some of the same sorts of things that contribute to test-score variability. ERP/IQ correlations do not necessarily say more about the biological basis of IQ than test scores, because ERPs are not necessarily more "biological" than other sorts of behavior. Indeed, from about 100 msec on following a stimulus, ERP components are better explained using behavioral terms such as those used in explaining reaction times than by reference to neural processes, as reflected in brainstem evoked potentials. A number of illustrations come to mind, but I will give one from work we have been involved with.

Schechter and Callaway (1984) used displays of large letters constructed from small letters as described by Kinchla (1974). Letters F, H, and Z were used to generate the 9 possible letter–letter combinations. There were three tasks. One was to respond only when large Zs appeared, one was to respond only to small Zs, and the third was to respond to any Z, large or small. P3s to big Zs made of small Zs had shorter latencies when the subject was attending only to large Zs, and longer latencies when the subject was attending only to small Zs. Thus, given the same stimulus, P3 latency varies as a function of the subject's strategy, just as is often the case with RTs in more conventional tasks (Hunt 1980).

That is not to say that ERPs may not help in locating the processes that account for *g*. However, we will need to manipulate the processes underlying both ERPs and test performance more precisely instead of relying on weak correlations. I will illustrate what I mean by suggesting an experiment that uses ERPs. Since stimulus complexity slows both P3 and RT, while response complexity slows RT without slowing P3, we can get some idea about what processes are involved by seeing how much an independent variable influences P3 as it changes RT. Thus, the stimulant methylphenidate can speed RT without changing P3, so we infer that it may act post-P3 and largely on response-related processes (Callaway 1984). On the other hand, the anticholinergic drug scopolamine can slow RT and P3 almost equally (Callaway, Halliday, Naylor & Schechter, in press). This suggests that scopolamine may slow pre-P3 stimulus-related processes. Now, suppose a low-*g* group had P3 latencies and RTs that were both slower than those of a high-*g* group and by the same amount. That would suggest that *g* is a function of stimulus-related processes. If, as is the case when one compares young and old groups, P3 in the low-*g* group was slowed by about half as much as RT was slowed, then we would suspect that the processes determining *g* are involved in response selection and execution as well as in stimulus evaluation.

I like the idea that *g* has to do with information processing, and there is certainly a lot we don't know about the biology of human information processing. From the psychological side, Hunt (1980) has made some progress in trying to determine which processes are related to *g* and which are not. I think ERPs could be used as I suggested above, and I think the additive factor method (Sanders 1983) might also be useful in isolating psychological processes involved in *g*. Then, to investigate the biological bases for the various information-processing operations of interest, one must be able to manipulate biological variables. Modern psychopharmacology offers a dazzling array of tools for manipulating biological variables, and cognitive psychologists are just now beginning to use them for their own purposes. There seem to be two styles in science. One is to be bright enough to perceive the truth quickly in nature. Research then consists of looking for examples to help the less gifted see the light. Then there are those who (perhaps for good reason) are more humble and at least aspire to the ideals described by Platt (1964). They are likely to be more intrigued by what they don't know than by what has already been revealed. There are even some bright people who have found this second style of science rewarding. If Jensen would like to shift from demonstration to investigation, I believe he will find that drugs, age, and certain diseases will serve as more useful biological variables for studying brain function than will skin color.

## The issue of *g*: Some relevant questions

Jerry S. Carlson

*School of Education, University of California, Riverside, Calif. 92521*

In his assessment of Spearman's hypothesis Jensen provides a valuable and scholarly review of research and theory concerning *g*. His arguments and conclusions are strengthened by the fact that the data are drawn from sources representing divergent research paradigms. Jensen's essay not only is informative, but presents a significant challenge: Neither our understanding of the nature of *g* nor our knowledge of the reasons for what appear to be reliable between or within group differences in the abilities involved is complete or even satisfactory. Accordingly, sufficient explanation of the differences cited cannot be made at this time; several research questions must first be answered. I would like to suggest just three, offering summary commentary with each.

1. How universal is *g*? The universality of basic cognitive abilities continues to present a challenge to cross-cultural psychologists. Although it is a fact that we are all of the same species, sharing certain necessary biological, linguistic, social, and cultural characteristics, does this ipso facto imply panhuman abilities and competences? From my reading of the cross-cultural literature, there seems to be reasonable evidence to conclude that our similarities (and this includes basic mental abilities) far outweigh our differences, although the latter most often seem to gain our attention. Wober (1974) has shown that non-Westernized Africans view and value what they have ecologically defined as intelligence in ways different from their literate, Westernized counterparts. But even minimal schooling and enculturation tend to recast previous conceptions of intelligence to conform more or less to Western definitions. The most recent work of Dasen (1984) is informative on this issue. He demonstrated that among the Baoulé of the Ivory Coast,

enculturation changed the traditional view of intelligence in dramatic ways. The change was not only linguistic and definitional but operational as well, evidenced by the positive correlations obtained between independent rankings of children on Piagetian and memory tasks with those made by literate adults but not those made by illiterate adults, all of whom knew the children and were asked to rank them according to the adults' definition of intelligence. This investigation, as well as other recent studies, suggests that *basic* cognitive competences may be more or less universal and may potentially involve the ability Jensen calls *g*. The ecological significance of these competences can vary within any society, of course, but as modernization occurs they may increase in importance. Research to investigate these issues would be useful from both theoretical and practical perspectives.

2. How modifiable are purported measures of *g*? Research designed to ascertain the modifiability of performance on *g*-loaded measures can make a significant contribution to our understanding of the factors, other than what Jensen terms the "hardware," that are involved in within- and between-group variability in *g*. Several lines of research have shown quite conclusively that modifications in testing approach and procedures can lead to improved estimates of cognitive competence for both individuals and groups and that performance measures, such as those cited by Jensen, may provide inadequate estimates of ability. Some of our work, for example, has shown that requiring individuals simply to describe verbally the task at hand and their approach in solving it can lead to significant improvements in performance on the most conceptually difficult items of the Raven matrices, the Cattell Culture Fair Test, and Piaget-derived tasks. Furthermore, we have shown that the reasons for improvement tend to be related to reduction of the negative or performance-diminishing effects of anxiety, impulsivity, and lack of motivation. (See Carlson & Wiedl 1980; Bethge, Carlson & Wiedl 1982.) The issue of whether or not test modifications can lead to substantial reduction in black–white differences on *g*-loaded tests is unclear at this point, however. There is some evidence that this may be so (Bridgeman & Buttram 1975; Dillon & Carlson 1978), but our most recent attempt (Carlson 1983) to replicate earlier findings have indicated that verbalization led to approximately equal gains on the Raven and Cattell tests by both black and white children.

Successful large-scale intervention projects, as represented by the work of Ramey and associates in North Carolina and Heber and Garber in Wisconsin (Heber & Garber 1973), are informative and potentially of great significance. The question of whether or not substantial changes in *g* can be brought about by the interventions is open; but the evidence is clear that important cognitive abilities of black and underprivileged youngsters can be improved. There seems to be hope that the differences Jensen reports for g may at least be reduced if extension of the efficacious treatments can be made to include large numbers of individuals.

3. What are the relationships between reaction times, evoked potentials, and *g*? Although the research paradigm that involves reaction time and other putative measures of physiological response has a tradition that goes back to Galton, has there been regeneration of interest in this area only recently. Eysenck and his associates have reported truly remarkable correlations between average evoked potential (AEP) within 256 msec of stimulus onset, elicted by auditory stimuli of 85 decibels, and IQ. The correlations cited are generally greater than 0.70. Unfortunately, to my knowledge, no thorough and independent replication of this work has been done. This will be critical before Eysenck's results can be confidently evaluated. Beyond this is the problem concerning the veracity of the AEP itself as an index of some neural substratum of intelligence. Evoked potentials are complicated and relate to different events. Early evoked potentials, of the sort reported by Eysenck, indicate anatomic development of the cortex and the fact that pathways to the cortex are functioning, but they do not indicate the degree to which the stimulus is processed by the peripheral receptors, more central nuclei, or the cortex (Parmellee & Sigman 1983). Measures of these more central sorts of phenomena may best be made using the event-related potential paradigm and focusing on waves beyond 256 msec. A further caveat comes from the fact that only stimuli of 85 decibels will apparently elicit the pattern of waves (amplitude and variability) that yields the high correlations with IQ measures. Is this because of some specific, undefined exogenous factor? This plus several other questions must be answered before conclusions can be made concerning the relation between AEP and intelligence.

The reaction-time (RT) data Jensen presents are interesting and provocative. The reliability of the correlations between reaction-time parameters and IQ is impressive, although alternative interpretations may be made concerning some of the relationships. One central problem concerning the Hick paradigm is that although correlations between *g* and RT are expected to increase across bits of information, most investigations do not show a clear trend in this direction. The most consistent correlation, on the other hand, is between the standard deviation of RT (intraindividual variability) and intelligence. The reasons for this are unclear at this point. We (Carlson, Jensen & Widaman 1983) have shown that voluntary, sustained attention may be involved in the relationship, but so may other factors such as arousal and orientation. Substantial research is required before we will be able to understand the functional relations between reaction-time measures and psychometric intelligence and the implications these have for individual and group differences in *g*.

## Different approaches to individual differences

Thomas H. Carr and Janet L. McDonald
*Department of Psychology, Michigan State University, East Lansing, Mich. 48824*

Jensen argues three main points: (1) Group differences between blacks and whites on current IQ tests are mainly accounted for by Spearman's *g*. (2) Differences in *g* are partly due to differences in the speed and variability of elementary mental operations. (3) Therefore, differences between blacks and whites on IQ tests are partly due to differences in the speed and variability of elementary mental operations.

Points 2 and 3 are attempts to use underlying cognitive processes to explain *g*, and for that reason they may be more interesting to cognitive psychologists than the first point. Without such attempts *g* remains, from the cognitive point of view, a substantially meaningless construct – statistical rather than theoretical and predictive rather than explanatory. Let's do the arithmetic necessary to quantify the argument, then examine some differences between Jensen's approach to cognitive analysis of individual differences and others.

Taking correlations reported in the target article and squaring them, it appears that somewhere between 12.9% and 72.3% of the variance in black–white differences on IQ tests is associated with *g*. In turn, 16.8% to 36.0% of the variance in *g* is associated with a general speed factor extracted from performance on a battery of reaction-time tasks. Hence somewhere between 2.2% and 26.0% of the variance in black–white differences on IQ tests is associated with both g and the general speed factor.

Psychologists never turn down the opportunity to account for 2–26% of the variance. It is clear, though, that knowing something about speed of processing (or *g*, for that matter, given the range of estimates) still leaves a lot to know about black–white differences on IQ tests. It is also clear that, at least if one sticks to the route of identifying variance held in common by all links of

the chain, knowing something about speed of processing leaves even more to know about differences in those things IQ tests are supposed to predict to begin with: school achievement and, secondarily, career success.

Nevertheless, the target article reveals that there is a relationship worth investigating between some general factor in test performance and some general factor in speed of reaction-time performance. But what could that relationship be? Here it might be useful to compare Jensen's psychometric approach with an approach called componential or component skills analysis.

Whether one examines work on intelligence (e.g., Carr 1984; R. Sternberg 1984b), reading comprehension (e.g., Carr 1981; Jackson & McClelland 1979; Olson, Kliegl, Davidson & Foltz 1984), mathematical computation (e.g., Ashcraft & Stazyk 1981; Groen & Parkman 1972; Klahr & Wallace 1976), expert problem solving (Chase & Simon 1973; Engle & Bukstel 1978), memory judgments (S. Sternberg 1969), or perceptual recognition (e.g., Allport 1980; Carr, Pollatsek & Posner 1981), one finds cognitive psychologists dividing the labor involved in a given performance into parcels that can be handed over to specialized processing mechanisms whose job is to carry out one particular kind of mental labor on some particular class of stimulus inputs. These specialized processors, or elementary mental operations, become the building blocks from which the performance as a whole is pieced together in a way that is somewhat analogous to piecing together commands and subroutines into a computer program (Posner & McLeod 1982). The goals of this enterprise are to identify the set of mental operations that is involved in any given performance (and in the course of looking at many performances to establish the repertoire of mental operations available for all performances), to identify the organization and patterns of information flow of the system set up from these operations to accomplish the performance, to determine the means by which the system is controlled and its component mental operations coordinated to achieve the performance, and finally, to identify parameters of mental operations or the system they comprise whose variation is responsible for individual and developmental differences in the system's overall effectiveness and efficiency.

These goals lead component skills analysts to do several things differently from Jensen. First, tasks are chosen to expose particular mental operations. Because no task recruits only a single operation for its performance, this is difficult. Three major strategies have been taken in the literature. The first involves comparing performance on two tasks that, on logical analysis, would seem to differ by a single operation: One task depends on a specific sequence of operations and the other depends on that sequence plus one more. This is the "subtractive" technique of Donders (1868/69; 1969). If the difference in performance between the two tasks varies with a dimension of between-subject individual differences, then the individual difference is attributed (at least in part) to that operation.

The second strategy again begins with logical analysis of a task into a sequence of mental operations. After the analysis, one identifies for each putative operation a stimulus manipulation that should (again, logically) influence that operation but not the others, one verifies that this is the case by showing that the manipulations do not interact with one another, and then one looks for interactions between each of the operation-specific stimulus manipulations and a dimension of individual difference. The individual difference is attributed to mental operations whose diagnostic stimulus manipulations interact with the subject variable. This second strategy is an application of S. Sternberg's (1969) "additive factors" technique.

The third strategy depends upon empirical identification of tasks for which variation in performance mainly reflects variation in the processing done by a single mental operation, even though the total performance may involve a much larger number of operations. This has been called the "isolable subsystems" technique (Posner 1978), and an example is the attempt by Carr et al. (1981) to establish physical same–different matching as a model task to study visual code formation in word recognition. If a battery of tasks can be constructed so that each of the component tasks primarily reflects a different mental operation or group of operations, then the relative strength of correlations between performance on the various component tasks and performance on a criterion task representing the entire performance will indicate which operations contribute most to individual differences in the performance.

Most component skills analyses use a mixture of the three strategies. In all such analyses, however, individual differences are pursued in one or both of two ways. The first is to try to identify particular mental operations whose characteristics distinguish individuals from one another. The distinguishing characteristics may be speed, accuracy, variability, capacity demands, or degree of sensitivity to various stimulus properties, and the assumption is made that not all mental operations will figure in the explanation – some will have characteristics that correlate with the individual difference of interest and some will not (e.g., Carr 1984; Frederiksen 1980; Jackson & McClelland 1979).

The second way of pursuing individual differences is to seek characteristics of the system, rather than characteristics of particular operations, that correlate with overall performance. In this case the assumption is that two people may possess identical repertoires of mental operations yet differ in performance because the operations are organized differently or exchange information with one another according to a different set of rules. This possibility has led cognitive psychologists to investigate interactions between mental operations (e.g., Omanson 1985; Schwartz & Stanovich 1981; Stanovich, West & Feeman 1981) and to examine patterns of intercorrelation among component tasks (e.g., Carr, Brown & Vavrus 1985; Evans & Carr 1985; Guthrie 1973; Olson, Kliegl, Davidson & Foltz 1984).

Neither of these approaches is quite the same as Jensen's. Roughly, Jensen appears to be seeking an operating characteristic that distinguishes *all* the mental operations of one individual from those of another, or perhaps the characteristic that, across the repertoire of operations, correlates most often with the overall performance in question. Such a search for the universal or the modal distinguishing characteristic is ambitious. Note, though, that Jensen is searching for this characteristic in a sample of tasks pulled unsystematically from a grab-bag rather than choosing tasks on the basis of one or more of the theoretical strategies taken in component skills analysis. In addition, he is focusing from the outset on speed and variability as possible operating characteristics, ignoring other possibilities such as capacity demands (cf. Carr 1984) or sensitivity to various stimulus properties (cf. Stanovich & West 1979). Finally, he is focusing on the characteristics of mental operations to the exclusion of characteristics of the system and its organization (cf. Vavrus, Brown & Carr 1983; Carr, Brown & Vavrus 1985).

Beyond these theoretical concerns, Jensen has adopted a methodology that makes interpretation of the speeds and variabilities that he observes problematic. The subject holds down a central button at the beginning of each trial of each choice reaction-time task. When the choice stimulus appears, the subject releases the button and presses one of the response buttons arrayed around it to indicate his or her choice. The latency attributed to the choice is the time from stimulus onset to release of the central button. Jensen apparently assumes that subjects release the central button at the moment they complete the choice decision. This need not be the case. In fact, clever subjects intent on maximizing speed and accuracy would treat onset of the choice stimulus as a simple reaction-time task, releasing the button immediately, then make the choice in a leisurely and reflective fashion, carefully monitoring themselves for accuracy. Given what is known about how simple and choice reaction-time tasks differ in speed, in variability, and in

susceptibility to interference from concurrent activities, all of Jensen's results would follow if subjects who differed in *g* also differed in their tendency to adopt this maximization strategy – and such an explanation would work even if subjects did not differ at all in the speeds of their elementary mental operations.

All of this is not to say that Jensen's approach is a poor one, though it has some flaws. He has made an important contribution toward a process explanation of a psychometric construct that has traditionally attracted considerable interest. His hypothesis, that speed and variability in the repertoire of elementary mental operations is a major determinant of *g*, ought to be energetically explored. However, we believe that such exploration will be more rigorous, more systematic, less susceptible to errors of omission, and more likely to produce a coherent and defensible set of findings if its theoretical and methodological underpinnings are expanded along the lines that have been followed in component skills analysis.

## Intelligence and *g*: An imaginative treatment of unimaginative data

Raymond B. Cattell

*Professor Emeritus, University of Illinois; 622 Kalanipuu Street, Honolulu, Hawaii 96825*

Jensen sets out, with impeccable scientific method, to supply the first possible alternative corroboration one would want to see, to his finding of significant intelligence differences between blacks and whites. The corroboration is that the kinds of tests on which there exist the greatest mean black–white differences are systematically those found to have the higher *g* saturation (correlation with the general intelligence factor).

Others will doubtless find various matters for comment in the rich array of data Jensen analyzes, so I shall confine myself to a single shaky step in his conclusion – namely, his use of *g* as the operational measure of intelligence.

It has been known dimly since 1940 (Cattell 1940), and with considerable precision since the sixties (Cattell 1963; Cattell 1967; Horn 1965; 1966) that Spearman's *g* actually factors into *two* main factors, $g_f$, *fluid intelligence*, and $g_c$, *crystallized intelligence*, which differ considerably in such matters as the tests of highest loading, the life course plots, the degree of inheritance, the reaction to brain injury, and the size of standard deviation of IQ.

Most of us devoid of prejudice have been inclined to interpret Jensen's black–white differences as largely differences in $g_f$, which is highly heritable. But Spearman's *g* – even when the array of cognitive tests of which it is the first component does meet the tetrad difference criterion – is actually a mixture of $g_f$ and $g_c$. The *investment theory* of intelligence asserts that $g_c$ appears as a unitary factor through the investment of $g_f$ in school and general culture, and its relative variance in the combined *g* measurement is a function of the relative genetic and cultural variance in the given group (not that $g_f$ is immune to *physiological* variance, e.g., in the early environment).

Jensen had to base his conclusions on such traditional tests as the WAIS and the WISC, which measure mixtures of $g_f$ and $g_c$, because psychologists have been slow in shifting to the Culture Fair, $g_f$, measures (Cattell 1940). His study should have been based on Culture Fair tests, like those of IPAT (1950; 1959) or the Raven matrices, but the abundant samples he needs were not available yet in those instruments. His conclusions must therefore be considered to be within the limitations of available last-generation data.

As if to strengthen the view that he is dealing with a more innate factor based more on laboratory, physiological measures than on pencil-and-paper behavior, the author turns in the latter part of his article to recent reaction-time studies and similar evidence for a conception of intelligence in line with the computer as "information-processing capacity." Despite the recent emphasis on reaction-time and brain-wave data by Eysenck (1982a), and Ertl (1969), these findings have long been put in perspective (Horn 1968) as lesser manifestations of $g_f$ correlations that do not necessarily make the estimated $g_f$ more "physiological and innate" than the pencil-and-paper measures of $g_f$ or $g_c$. Intelligence is not, in any satisfactory definition, "information-processing capacity" (with its suggestion of computer science) but, in essence, "capacity to perceive relations." These perceived relations are either more innately fixed in $g_f$ or more learned in $g_c$. The computer is a false model for intelligence, because its construction is basically different, with all-or-nothing discharge of units and the absence of effects from the electromagnetic field, shown to be active in the brain. It is interesting, however, to see how the black–white difference extends into these laboratory performances.

According to Jensen's target article we can draw the conclusion that a substantial black–white difference in *g* is corroborated by the factor loading order. But since we do not know exactly what percentages of $g_f$ and $g_c$ enter into this *g* we cannot, at more than a probabilistic level, conclude that the black deficiency is in the more innate $g_f$ or the more environmental $g_c$. The latest figures for heritability, in terms of *interfamily* differences, are, for $g_f$, 89; and for $g_c$, 29 (Cattell 1982, p. 312). Allowance for error would probably raise these values somewhat, but it is clear that the debates over "the inheritance of intelligence" have quoted different figures through using tests, like the WAIS and WISC, that are undefined mixtures of two distinct factors.

Jensen is not unaware of this point, but he hopes to avoid the difficulty by saying that "the correlation between $g_c$ and $g_f$ . . . [is] so high, in fact, that these two facets of general intelligence cannot always be clearly distinguished. . . ." The most careful rotational studies give correlations, actually, of .47 in 14-year-olds and .18 in general adults (Cattell 1971, p. 96). This means that a quarter or less of the variance of crystallized intelligence is due to fluid intelligence – consistent with the low value for heredity, .29, found for crystallized intelligence.

Despite the neatness and thoroughness of Jensen's check on Spearman's hypothesis, we are left with results still arrested at Spearman's first concept of *g* (1904). Although I would tend to conclude, from other evidence (Horn 1968), that much of the black–white difference is located in $g_f$, the present evidence leaves this only as a probability. MacArthur and Elley's (1963) study on 271 children found the saturation of various tests with the *g* factor (defined by the sum of all) to be as in Table 1.

These results not only agree with the saturation order obtained by Jensen but also show that $g_f$ measures (IPAT Culture Fair and Raven) rank very high – so high, in fact, that we may perhaps best consider *g* to be more $g_f$ than $g_c$. Thus, at a rougher practical level we may consider Jensen's check on Spearman's theory to apply more to $g_f$ than to $g_c$ and to imply a more genetic component in the black–white difference.

As another practical, social conclusion from Jensen's analysis

Table 1 (Cattell). *The g saturations of some common cognitive ability tests*

| | |
|---|---|
| IPAT Culture Fair (Scale 2A) | .75 |
| Raven matrices | .71 |
| Large Thorndike Number Series | .55 |
| Reading vocabulary | .34 |
| Reading comprehension | .50 |
| Arithmetic reasoning | .46 |
| Spelling | .20 |

*Note:* The pool taken to estimate *g* was larger than this sample of tests, from MacArthur and Elley (1963).

one would like to see his Table 6 extended to show the relative numbers of blacks and whites in these occupations. In spite of "fair employment" injunctions the percentage of blacks should, to fit Jensen's conclusions, be inversely correlated with the mean intelligence levels of persons holding the occupations (Cattell 1971, p. 451). Results on this could easily be obtained.

In relation to possible ultimate conclusions, psychological and social, one cannot help regretting, in a sense, that such intelligent and thorough analysis, accompanied by quite unusual statistical finesse, has had to be lavished on Spearman's primitive (1904) theory of *g*. But such are the lags of scientific thought that even if Jensen had couched his questions in terms of the newer known structures of $g_f$ and $g_c$ he would not have found, yet, enough data in the literature to work upon. In short, this is as valuable a contribution, clearly supporting a hypothesis, as the present field of data will support.

## Interpretations for a class on minority assessment

J. P. Das

*Centre for the Study of Mental Retardation, University of Alberta, Edmonton, Alberta Canada T6G 2E1*

If I were teaching a course on minority assessment and my students were mainly blacks and sympathetic whites who believed that blacks have been victimized in American society, I would have a problem in getting the class to accept the target article's contents and its implications. I would perhaps prepare the following lesson plan.

I shall point out to my students that the paper starts with assuming the gist of Jensen's 1969 paper in the *Harvard Educational Review* as accepted fact. This paper ignores the existing valid criticisms that question the meaning of the difference between black and white IQs and the mechanism by which the difference was established. There is also a tacit acceptance of the Level I–Level II distinction, which has been rejected for various reasons, including that all cognitive activities cannot be contained within this dichotomous division (Jarman 1978) and that it is simplistic (Cronbach 1969).

Then the paper quickly moves on to Spearman and *g* and the "discovery" that black–white differences may essentially reflect the differences in *g*. At this point, I will cite Cronbach's (1969) comment on Jensen's treatment of *g*: "Jensen protests that we should not 'reify *g* as an entity,' but it seems to me that he does so especially as he begins to insist that it is a 'biological entity'" (p. 197). Anyway, my students will have gotten the impression that Jensen is confirmed in the belief of *g* being a reality, that he regards it as something like a cosmic spirit that seeks manifestation, in a polymorphous manner, in all human behaviors that do not short-circuit the cortex.

Next, we will learn about "elementary cognitive tasks," or ECTs. How is cognitiion reflected in the elementary reaction-time (RT) task used by Jensen, which he admits has little intellectual content? The central problem in cognition, according to many, is to understand how knowledge is represented in memory. So how can the RT task of Jensen (his Figure 8) fail to reflect stimulus preprocessing, stimulus categorization, response selection, and response execution, the basic components of RT described in textbooks (Lachman, Lachman & Butterfield 1979) but at the same time be a cognitive task?

The RT task entails many different time measures: the time for the initiation of the ERP (event-related potential), time between the ERP and onset of the electromyogram (EMG), then the time between EMG and response initiation.

At this stage, my students will impatiently ask which one of these is indicative of "information-processing speed" as used by Jensen. They will be advised to wait until later, and asked to consider the theory in terms of "working memory" as a basis for individual differences in *g*. Even if every student believes in the explanation offered, the concept of working memory itself is being reexamined in contemporary psychology (Klapp et al. 1983).

We will now have reached the section "Methodological desiderata" in Jensen's paper, and the black–white issue is quite explicit at this point. The students will learn that inequalities in intelligence between the two populations are not to be erroneously attributed to cultural or linguistic factors. The "population differences" in ability are valid. I will pacify my class at this point by quoting Cronbach (1969), who wrote that at times striking differences in "ability" can be overcome very simply.

The class will then read the next few pages – a statistical teach-in, decontextualized from the background of strong emotions raised in the preceding statements of Jensen's paper, until the class confronts the chronometric studies. These chronometric studies use tasks that have very little "intellectual content" but correlate positively with tests, often scholastic tasks, that are filled with intellectual content and require a specific knowledge base! Is it logical, then, to assume that what they have in common cannot be intellectual skills, but factors which are extraintellectual, which can then be manipulated to bring up performance? My class knows the disadvantages of growing up as a black person, the deprivations that breed apathy, create self-doubts and lower one's self-esteem, so that the black testee may not acquire the appropriate attitudes and motivations for taking chronometric tasks in the environment of the laboratory.

But I will bring back the class to a scientific study of the task itself, in order to determine the best correlate of *g*. Which components of the task are likely to reflect *g* if we analyze an RT task such as Saul Sternberg's memory scanning? The example is an experiment Karrer (1984) did on mentally retarded adolescents (low IQ), comparing their performance with normal adolescents of the same age and with younger children of the same mental age. He, like Jensen, used a home button, but also two others, one on each side of the home button; subjects were to hit one of these to indicate "Yes" and the other for "No." The most interesting part of his study is the examination of the *return time*, after the response had been executed, and the *central time*, which should be sensitive to *g* differences (IQ) and to the information load in Sternberg's task. The return time should not reflect IQ difference. However, if this were not the case, then we should rethink "information-processing speed" and how useful it would be in generating testable hypotheses concerning *g*. The findings were as follows: Central time was longer when subjects searched for five items than for one; it varied with information load. But there was no difference between the mentally retarded and normal mental-age-matched children. On the other hand, return time for the retarded group was longer in the five-item than in the one-item task, and what is strange, the retarded took the longest time to return to the home button – longer than normal adolescents.

The class will probably experience information overload at this point and wish me to end the lesson. We will come to the future of this line of research. It is harmless if one is curious to know about the relationship between statistical facts and artifacts. It is harmful if blacks are declared slow in information processing on the basis of this paper. Blacks in America have surpassed everyone else in speed and the judicious use of that speed in dancing and athletics, to take only two instances. The Olympics are still fresh in the memory of my class. Should we spend American resources, intellectual and financial, to support the antiquated hunch of a British professor about the inferiority of American blacks?

## The nature of cognitive differences between blacks and whites

H. J. Eysenck

*Department of Psychology, Institute of Psychiatry, London SE5 8AF, England*

I have recently surveyed "the effect of race on human abilities and mental test scores" (Eysenck 1984), and the major conclusion of this survey was that there are very marked population differences in IQ test scores. There is a general decline of IQ mean scores, ranging from the Mongoloid peoples, particularly the Chinese and the Japanese, through Northern European Caucasoids and their descendants, to Southern European Caucasoids and Indians, to Malays and Negroid groups. In each group, of course, there may be and frequently are differences between one subgroup and another; thus, within a given Caucasoid group the Jews usually have an unusually high mean IQ as compared with non-Jews. It has also become apparent that there is a close correlation between the IQ level of given groups, their socioeconomic status, and their degree of cultural achievement. These generalizations are based on direct empirical findings, but of course their interpretation is not immediately obvious. In particular, it has been questioned to what extent IQ tests are measures of intelligence, and a debate has been raging about environmental or genetic causes of the observed differences.

With respect to the meaning of the term "intelligence," there has been a long-standing debate between the followers of Sir Francis Galton and those of Alfred Binet. For Galton, intelligence was a largely innate property of the central nervous system and the cortex in particular, predisposing a person to be proficient or otherwise at any test of cognitive skill, such as problem solving, learning, remembering, organizing, or following directions. For Binet, intelligence was largely an artifact, the average of a number of independent abilities, each of which was subject to educational, cultural, and other environmental influences. This different understanding of the term "intelligence" has played havoc with the debates that have taken place among psychologists in an effort to arrive at a satisfactory solution to the problem. Clearly we are dealing with three different conceptions, which have often been called Intelligence A, Intelligence B, and Intelligence C.

Figure 1 illustrates the relationship of these three concepts. Intelligence A embodies the central meaning of Galton's conception; it is the largely or entirely innate capacity of the central nervous system and cortex to process information correctly and without error. Intelligence B embodies Binet's concept of "social intelligence," that is, a person's capacity to use Intelligence A in a great variety of social situations. Intelligence B is much more inclusive than Intelligence A, because it involves many additional factors, such as personality, education, cultural influences, and socioeconomic determinants; and it relates to a host of different cognitive performances, such as comprehension, memory, learning, problem solving, judgment, reasoning, adaptation to the environment, and the elaboration of strategies. Intelligence B is more like the popular conception of intelligence, but of course it has no scientific status, being a compound of many different influences, of which Intelligence A is only one.

IQ is positioned between these two, being more inclusive than Intelligence A (because obviously the tests used incorporate cultural and educational material, and because personality qualities, such as anxiety, cannot easily be separated from ability). IQ is related to Intelligence B because it clearly has great social implications, as indicated by the high correlations between IQ and educational success and life success in general (Eysenck 1979.) The close relationship between IQ and Intelligence B has misled many students of the field in recent years to deny the existence or importance of Intelligence A or its relationship with IQ (Sternberg 1982). Eysenck (1985) has recently surveyed the evidence from a group of psychophysiological studies of intelligence indicating that we are now beginning to have some idea of the biological nature of Intelligence A, and suggesting that in essence Galton was right and Binet wrong in their assumptions about the nature of intelligence.

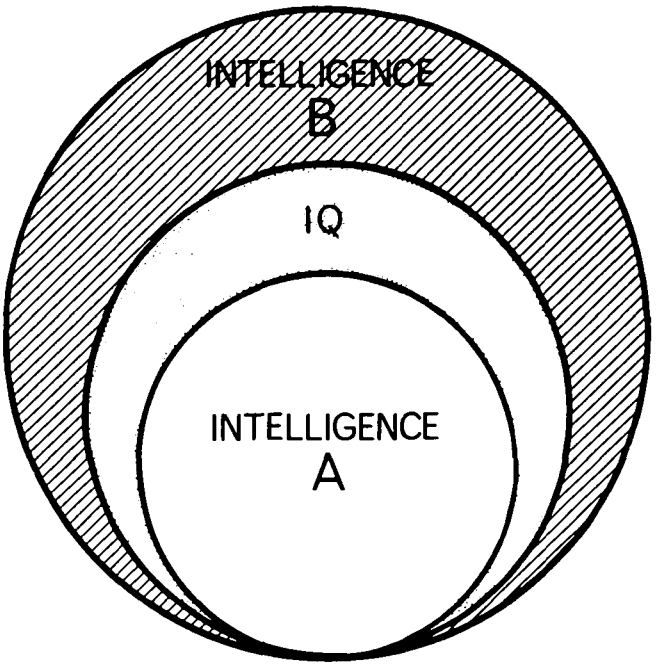


Figure 1 (Eysenck). Relative relations between Intelligence A, Intelligence B, and IQ.

These recent studies, indicating a close relationship between certain measures taken on the average evoked potential and IQ as measured by standard modern tests, give results that are quite impossible to reconcile with Binet-type notions about intelligence being an artifact and a mere statistical average without psychological meaning. This point is vital to any understanding and appreciation of Jensen's argument, which rests essentially on the recognition of Spearman's *g* as a fundamental psychological variable. Given this admission, Jensen's argument about what he calls "Spearman's hypothesis" seems to be irrefutable. The highly significant concordance between factor loadings on *g* and black–white differences would be difficult to explain on any other grounds than those used by Spearman and Jensen in putting forward their hypotheses. The observed correlations are of course much reduced because of the lack of range; inclusion of tests having only minimal relation to intelligence would almost certainly increase the size of the observed correlations drastically. The logic of the argument seems faultless; it depends entirely on the admission that recent evidence supports very strongly the existence of "intelligence" as a separate entity, measurable by IQ tests (although not perfectly) and relevant to social activities. Jensen's own work on the relationships between chronometry and intelligence lend further support to Galton's original conceptions, and it is now very difficult to doubt that in essence he was right. Granted this, Jensen's new step seems entirely convincing.

## The black–white factor is *g*

Robert A. Gordon

*Department of Sociology, Johns Hopkins University, Baltimore, Md. 21218*

***A classic test of a classic hypothesis.*** By establishing major three-way connections among the psychometric general (*g*)

factors of blacks and whites and mean differences between the two populations, incidentally vindicating claims that factor analysis can serve as a source of fertile theoretical constructs, Jensen has substantially enhanced the ontic status of the *g* construct. That all of these accomplishments redound to the credit of Spearman, at the center of controversy throughout his own career, is indeed impressive.

But Jensen has also attempted to go beyond Spearman (a) by linking individual differences in psychometric *g* to individual differences in reaction-time parameters and to the general factor of those parameters, (b) by implicating the complexity (manifest *g*-loadedness) of both psychometric and chronometric tasks in the degree to which the association between the two categories of differences holds, and (c) by extending the Spearman hypothesis to black–white (and other group) differences in reaction time, that is, by relating the magnitudes of those group differences to the mean reaction times of a series of simple tasks and, in turn, relating the mean reaction times of the tasks to their loadings on a psychometric *g* factor.

Whether or not the still novel reaction-time findings hold up over time, I agree with Jensen that his psychometric linkages alone effectively strip away the basis for contending that *g* depends greatly on individual differences in mastering specific information. Nice convergent-and-discriminant touches appear in Jensen's demonstrations that the Spearman hypothesis fails to account for mean differences between congenitally deaf and hearing individuals and that no other factor besides psychometric *g* is related to the *g* factor of the reaction-time parameters. Jensen's observation that the regression line in his Figure 1 passes through the true origin of loadings and black–white differences is another sign of unusual consistency in the results. A likely reason for that consistency is given in the next section.

***Perfect correlations may be sufficient for attributing the black–white difference to g, but are they necessary?*** For convenience, let us consider just the Pearson correlations that the white loadings produce for the Spearman hypothesis in Jensen's 11 batteries. The individual correlations range from .36 to .78, and the mean is .61. As Jensen makes clear, such correlations are subject to influences that would usually reduce them: restriction of range in loadings because all subtests measure *g* about equally well, suppression of the black–white difference because the variable also taps a factor on which the black–white difference is reversed (e.g., see Jensen's discussion of the motor coordination subtest in the Department of Labor battery), and anything else known to disturb data. Consequently, we have no developed standard, other than the usual ones for judging correlations, that tells us how to evaluate the outcome of a test of the Spearman hypothesis. Short of obtaining perfect or nearly perfect correlation, there is no way to know how large a nonzero correlation it is reasonable to demand as evidence.

Thus, Sandoval (1982) cautiously regarded a (rank) correlation of .48, which was significant with a one-tailed test, as not "strongly supportive" (p. 200) of the Spearman hypothesis. A number of Jensen's correlations are lower than .48, yet Jensen, correctly in my opinion, regards all of his sets of data as consistent with the hypothesis. Many readers may grant that Jensen's mean correlation of .61 is a nontrivial result yet still not know what attitude to adopt toward the residual black–white difference or what to make of the batteries that yielded lower correlations.

Clearly, there is a problem with using correlations alone to test the hypothesis. Correlations measure covariation with respect to variation around the local mean, no matter how trivial that variation may be. Indeed, it is virtually axiomatic that the better an intelligence battery has been constructed, the more difficult it will be to find evidence for the Spearman hypothesis. The axiom is borne out by Jensen's demonstration that correcting for attenuation (i.e., simulating perfectly reliable measures) in seven batteries reduces their correlations testing Spearman's hypothesis by about 11% (white loading), because it reduces variability between subtests (see also Jensen's remarks on the ASVAB test). The local mean serves as a merely conventional origin (zero point), and there is nearly always variation around it, but that variation can be modest in amount and elusive in its derivation. Consequently, although a correlation is suitable for assessing how much of that variation the Spearman hypothesis accounts for, the same correlation may be unsuitable for identifying the underlying nature of the black–white difference – unless, of course, the correlation approaches 1.0. Thus, the task of assessing variance needs to be distinguished from the task of identifying what construct the population difference represents, if any.

Mean black–white differences can be expressed as point-biserial correlations. Such correlations can be viewed as subtest loadings on a black–white population factor or component, and that factor can be compared with *g* via the same coefficient of factorial similarity (or congruence) that Jensen used to compare general factors of blacks and whites in his Table 3 (see his note 2 for the formula).

The factor similarity coefficient (Harman 1960, p. 257) measures covariation with respect to variation around zero, rather than around the local mean. That zero is a meaningful one on the absolute scale of values taken by correlations, hence comparisons based on the coefficient remain on the same absolute scale from one application to another and from one factor to another. They also remain sensitive to the scale on which the correlations of the original factored matrices were expressed and to the signs of those correlations as reflected in the signs of loadings. In contrast, variation about the mean loading need have no relation to the scale or signs of original correlations, and so it is easy to contrive extreme examples in which the correlation is −1.0 but in which the similarity coefficient is positive and virtually perfect. This final advantage concerning the scale of the similarity coefficient is reflected in the observation by Gorsuch (1974): "In the case of orthogonal components where the factor scores have means of zero and variances of one, the result of calculating coefficients of congruence on the factor pattern is identical to correlating the exact factor scores and is, indeed, a simplified formula for that correlation" (p. 253).

Although Jensen's general factors are not first principal components but principal factors, similarity coefficients reveal that they resemble the principal components so closely (e.g., Jensen & Reynolds 1982) that any coefficients of similarity based on them can be viewed as close approximations to the correlations between factor scores of components. In the orthogonal case, of course, the factor pattern mentioned by Gorsuch equals the factor structure.

By assuming that the unreported subtest standard deviations are equal in the black and white samples, it is possible to derive a standard deviation for both groups combined (see McNemar 1969, p. 24). With that combined standard deviation and the mean black–white differences in Jensen's Appendix, the differences can be expressed as point-biserial correlations (e.g., Guilford 1965, p. 322), and thus factor similarity coefficients can be used to supplement correlations in Jensen's tests of the Spearman hypothesis.

But first some details must be made explicit. I have assumed that black and white samples are equal in size in deriving the combined standard deviation, and I have also evaluated the point-biserial correlation for the case of equal samples. These decisions concerning sample size have virtually no effect on the resulting similarity coefficients. The more arbitrary assumption of equal standard deviations within both populations was evaluated against the actual standard deviations in the five batteries for which the original sources were at hand (Department of Defense 1982; Jensen & Reynolds 1982; Mercer 1984; Sandoval 1982; Scarr 1981). That assumption affected the similarity coefficient only in the third decimal place, and then by only two units at most. For disattenuated data, the point-biserial correlations were based on Jensen's disattenuated black–white differences,

Table 1 (Gordon). *Coefficients of factorial similarity between g factor loadings within each population and mean black–white differences expressed as point-biserial correlations in Jensen's 12 test batteries*

| Study | Uncorrected for attenuation | | Corrected for attenuation | |
|---|---|---|---|---|
| | White | Black | White | Black |
| Jensen & Reynolds (1982)[a] | .980 | .972 | .980 | .972 |
| Reynolds & Gutkin (1981)[a] | .970 | .960 | .969 | .960 |
| Sandoval (1982)[a] | .975 | .984 | .974 | .985 |
| Mercer (1984)[a] | .988 | .990 | .988 | .990 |
| National Longitudinal Study | .989 | .988 | — | — |
| Nichols (1972) | .964 | .962 | — | — |
| Dept. of Defense (1982) | .989 | .993 | .990 | .993 |
| Dept. of Labor (1970) | .915 | — | .916 | — |
| Kaufman & Kaufman (1983) | .943 | .940 | .941 | .936 |
| Veroff et al. (1971) | .960 | .960 | — | — |
| Hennessy & Merrifield (1976) | .975 | .979 | — | — |
| Scarr (1981b) | .963 | .977 | — | — |
| Mean: | .968 | .973 | .965 | .973 |
| Standard Deviation: | .022 | .016 | .027 | .022 |

[a]WISC-R study.

rather than on correcting the attenuated point-biserial correlation, as that seemed more faithful to his analyses.

Table 1 presents tests of the Spearman hypothesis based on factor similarity coefficients derived from the data in Jensen's Appendix. For good measure, I have included Scarr's (1981b) small battery.

Note first that the coefficients are not automatically all equally high. The Department of Labor's (1970) GATB yields the lowest values. However, if the GATB's factorially complex motor coordination test that Jensen himself remarked upon is excluded, the similarity coefficients for both populations rise to .958, a figure more in line with coefficients from other batteries. The second lowest coefficients belong to the K-ABC of Kaufman and Kaufman (1983). I have spent the past year analyzing the K-ABC and have found that it is not a univocal battery. In that respect, therefore, the K-ABC resembles the complex motor coordination test of the GATB – another exception that proves the rule. Even so, the K-ABC coefficients in Table 1, which are for the school-age sample, are not so low that one would reject the hypothesis that its general factor and the black–white factor are equivalent. But at five younger ages, with much smaller samples, I have found that the coefficients for the K-ABC deteriorate further, ranging from .63 to .81. Since coefficients below .46 have been rejected as evidence for factorial congruence, and since those of .94 (Harman 1960, p. 259), .90, or, more stringently, .95 (Jensen's note 2), have been interpreted as evidence of factorial identity, the coefficients for the K-ABC at the younger ages fall within a gray zone. Thus, the outcome of testing the Spearman hypothesis with the similarity coefficient is by no means a foregone conclusion.

The effect of the correction for attenuation on tests of the Spearman hypothesis illustrates how vulnerable the correlation is to even slight sources of variance. Take the two studies with the largest such effects, Sandoval's and Mercer's (see Jensen's Table 3). Within the two studies, the correction has such a slight effect that the white loadings correlate .99 and the black–white differences correlate .98 and .99, before and after correction. Yet the correction reduces by 8% and 9% the amount of variance in the between-population difference accounted for by *g*. As Table 1 shows, the coefficient of similarity, even in the Sandoval and Mercer studies, is hardly affected at all by the correction for attenuation.

According to the standards by which factors are usually equated, the average coefficients in Table 1 indicate that the black–white factor is *g*. This interpretation holds for all of the individual studies too, if one adopts the less stringent cutoff of .90. But the lower standard may not be needed, for even the marginal GATB is brought into line if its one problematic subtest is excluded, as I showed. Results for the WISC-R studies are especially strong and consistent, if one takes into account the reduction of the black–white difference in IQ due to socioeconomic matching in the Reynolds and Gutkin (1981) study. In light of Gorsuch's observation, factor scores based on *g* and on the magnitude of the black–white difference would correlate almost perfectly.

***Other indications of the reality and robustness of a latent trait such as g.*** Jensen's dense network of validation squares with other evidence of the fundamental reality of *g*. Evidence that some normally distributed latent trait may underlie IQ differences comes from examining backward digit span performance in six samples (four white, two black) ranging across a 46-year period, including blacks tested in 1918. When item passing rates in the six samples were transformed to unit normal deviates and then standardized for mean and variance, the mean absolute differences among all samples for corresponding items amounted to less than 1% when restated in terms of percentages passing (Gordon 1984).

This virtually perfect fit implies that the observed passing rates of the digits-backward items behaved as though they were ascending, descending, or straddling the hump of a normally distributed latent trait common to all of the samples. The absence of any significant group-by-item interaction contrasted markedly with the abundance of such interaction typical of authentic instances of cultural diffusion (Gordon 1984). Other evidence consistent with a one-standard-deviation difference between blacks and whites on a normally distributed latent trait (such as *g*) is reported in Gordon (1976) and Lamb (1983).

## Measuring and interpreting *g*

Jan-Eric Gustafsson

*Department of Education and Educational Research, University of Göteborg, S-431 26 Mölndal, Sweden*

Jensen's target article is careful in its interpretations and conclusions, but it goes without saying that if the origin of the black–white difference can indeed be localized to certain basic processes close to the "hardware" level of the cognitive system, this supports a theoretical framework that stresses genetic rather than cultural causes of the observed differences and that implies pessimism concerning the possibility of reducing the differences through social and educational interventions.

This commentary consists of three parts: first, the psychometric evidence referred to by Jensen is scrutinized; then the logic of investigations into elementary cognitive processes as a means of understanding the nature of *g* is commented upon; and finally alternative interpretations of the black–white difference are discussed.

***The psychometric evidence.*** The results presented by Jensen indicate that there is a correlation between a test's *g* loading and the magnitude of the standardized black–white difference. The relationship is far from perfect, however, and the interpretation of this result is not straightforward.

From a technical as well as a theoretical point of view, Jensen's approach to the study of Spearman's hypothesis suffers from the fundamental problem that the *g* factor is taken to be the dominating factor in the matrix of intercorrelations between tests, irrespective of which tests are represented in the battery. This implies that the estimate of the *g* loading for a test varies as a function of what other tests were included in the battery. More serious, however, is the fact that it also implies that the nature of the *g* factor is at the mercy of the composition of the test battery.

Evidence is accumulating (Gustafsson 1984) in favor of a particular hierarchical model of the structure of abilities, with Spearman's *g* at the apex of the hierarchy. This factor has, furthermore, been shown to be equivalent to the factor of fluid intelligence identified by Cattell (e.g., 1963) and Horn (e.g., 1968) (Gustafsson 1984; Undheim 1981b). Below *g* the model includes, among other broad factors, the second-order factors, crystallized intelligence (mainly school achievement and verbal competence) and general visualization (roughly competence in dealing with visual/spatial information), and at the lowest level the primary factors in the Thurstone and Guilford tradition.

The tasks most clearly related to *g* seem to be complex nonverbal reasoning problems that are new to the examinees, the Raven Progressive Matrices being the archetypical example. But such tests are infrequently represented in the test batteries of the studies upon which Jensen bases his analysis, and when included they tend not to obtain the highest *g* estimates. The tests most profusely represented in the studies are instead those measuring crystallized ability, and invariably those are the tests that come out with the highest *g* loadings.

The factor that Jensen interprets as *g* thus seems to be severely biased toward school achievement and the acquisition of culturally valued information and skills. In my opinion, therefore, Jensen's analysis leaves Spearman's hypothesis largely uninvestigated, and the hypothesis can neither be accepted nor rejected on the basis of the analyses performed.

It would carry us too far afield to discuss in this context alternative methods for investigating Spearman's hypothesis. However, a more appropriate method for simultaneously investigating the strong and weak forms of the hypothesis would probably be afforded by Sörbom's (1974) technique for analyzing differences in factor means.

***Elementary cognitive processes.*** The research on "elementary cognitive processes" through the reaction-time (RT) paradigm represents an attempt to reveal the psychological nature of the *g* factor, which only appears as a mathematical abstraction in the psychometric research.

To me the results reported by Jensen represent a most striking and elegant illustration of the role of task complexity in the elicitation of *g*. It would seem, however, that the effects of task complexity cannot be explained solely in terms of the additive effects of speed of execution of simple processes, so when Jensen says "that *g* essentially reflects the speed and efficiency with which a number of elementary cognitive processes can be executed," the emphasis in the interpretation should be on the *coordinated* execution of many processes. Coordination of processes is not a low-level process, however; it comes much closer to the concept of "metacomponents" than it does to the concept of "elementary cognitive processes." [See Sternberg: "Sketch of a Componential Subtheory of Human Intelligence" *BBS* 3(4) 1980 and "Toward a Triarchic Theory of Human Intelligence" *BBS* 7(2) 1984.]

The conclusions from the RT research and the psychometric research converge on the conclusion that an important characteristic of *g* is the ability to deal efficiently with complexity. While it is paradoxical that the rudimentary tasks employed in the RT paradigm so strongly enforce this conclusion, there is a strong need to take the further step within the RT paradigm, and others, to analyze the psychological nature of complexity. Until this is done there is little basis for understanding the *g* factor.

***Concluding remarks.*** In my opinion Jensen has not convincingly demonstrated the correctness of Spearman's hypothesis, and a firm conclusion will have to await results from a stronger analysis of pertinent data. However, the results, along with the research reviewed, indicate that at least the weak form of the hypothesis may eventually receive support.

Even though Jensen explicitly includes cultural influences as a possible explanation of mean differences in *g*, he seems to relegate them to a subordinate position, attributing limited influence to them, such as effects on reliability. But even though it is true that *g* reflects only those influences that are manifested in performance on all tasks, this does not preclude environmental explanations for observed differences.

Thus, psychometric *g* reflects variance from factors such as test-taking skill, persistence, and attitude, and these are certainly likely to come under strong cultural influence. Furthermore, to the extent that *g* reflects general problem-solving skills there is little reason a priori to assume heavier involvement of genetic factors in between-group differences in *g* than there is for more narrowly defined abilities.

Through intellectual heritage, perhaps, the *g* factor has come to be associated with characteristics such as immutability and strong genetic determination, which may be why it has been more or less banned from psychological research for several decades. Since the concept of general intelligence seems to be unavoidable, both in empirical and in theoretical research, it is reassuring that the last couple of years have brought a renewed interest in the concept. However, if this concept is to stay with us it is incumbent upon all of us to use it with utmost care and to be quite explicit about any assumptions we make.

## Do we know enough about *g* to be able to speak of black–white differences?

Ronald C. Johnson[a] and Craig T. Nagoshi[b]

*[a]Behavioral Biology Laboratory, University of Hawaii, Honolulu, Hawaii 96822 and [b]Institute for Behavioral Genetics, University of Colorado, Boulder, Colo. 80309*

In his commentary on Dr. Jensen's target article, Wilson (q.v.) summarizes the conclusion of our soon-to-be-published paper (Nagoshi et al., in press) that any group difference in *g* would of necessity be reflected in the tests that load on *g*. This finding in itself casts serious doubts on the validity of Jensen's conclusions concerning black–white differences in cognitive abilities. The present commentary is focused upon another issue that has arisen in the light of further analyses (to be formally presented in a forthcoming paper) of data from the Hawaii Family Study of Cognition.

In the introduction to his paper, Jensen notes a previous finding that on various subscales of the WISC the degree of inbreeding depression in the offspring of cousin versus non-cousin marriages is positively and significantly related to the *g* loadings of those subscales (Jensen 1983a), suggesting that *g* is more under the control of dominance genetic variance than are the non-*g* components of intelligence tests. Using the data from intact nuclear families of Caucasian and Japanese ancestries living in Hawaii and Koreans living in Korea, we have found that the degree of familiality (additive genetic variance plus common environment) across the 15 cognitive tests used in the Hawaii Family Study of Cognition was also positively related to the *g* loadings of those tests. The mean correlation of *g* loadings with parent–offspring correlations (all correlations reported here were corrected for test reliability) across the three ethnic groups and four parent–offspring combinations was found to be 0.54, while the mean correlation of *g* loadings with sibling correlations was found to be 0.42. On the other hand, *g* loadings were even more highly correlated with spouse correlations (mean $r$ =

0.63), which could account for the familiality results. In support of Jensen's emphasis on *g*, we found that tests that predicted parents' educational and occupational attainment were highly *g*-loaded (mean $r = 0.73$ for education, 0.67 for occupation). For Caucasian families, those tests for offspring that were most highly correlated with parental education, even after partialing out parental cognitive ability, were highly *g*-loaded (mean $r = 0.48$), while mean *r*'s for the other two groups and for the influence of parental occupation on offspring cognition were postive but near zero.

The above results suggest any number of plausible, untested alternative hypotheses to account for the ubiquity of *g* in these different cognitive ability relationships. The word to be emphasized here though is *untested.* Jensen is to be lauded for his extensive series of studies bringing attention to *g*, but there is clearly a need for even more basic research on the nature of *g*.

## Golly *g*: Interpreting Spearman's general factor

Lyle V. Jones

*L. L. Thurstone Psychometric Laboratory, University of North Carolina, Chapel Hill, N.C. 27514*

Arthur Jensen's reanalysis of data from 11 studies provides convincing evidence that the observed differences between average scores of black and white samples in the United States on a variety of mental tests are directly related to average differences in *g*, "the general factor common to all complex tests of mental ability." In view of the pervasive nature of the empirical findings that show average black–white test score differences, any other conclusion would be totally unexpected. For a wide variety of mental tests, average scores for one group of test takers are higher than average scores for another. A general factor, *g*, is defined to be that component which is common to the many tests, and the composition of *g* is found to be similar for both groups. It would be totally unexpected, then, to discover other than a direct relation between the tests' standardized group mean differences and those tests' loadings on *g*. (Such an unexpected result is reported by Jensen, where one population is made up of normal children and a second consists of preverbally deaf children: but, unlike findings for black children and white children, the composition of *g* – i.e., the relative loadings of subtests on *g* – is likely to be different for deaf children and hearing children.)

Jensen tells us that "the *g* factor is Spearman's label for the single largest . . . source of individual differences that is common to all mental tests," and notes that "Spearman's . . . theory of *g* as a kind of general 'mental energy' [is] of no particular relevance or importance in the present discussion. At this point, *g* need not be attributed any meaning beyond its operational definition in terms of factor analysis." In Jensen's analyses, *g* is always measured by a first principal factor or by a higher-order factor in a hierarchical solution. While the latter is preferable on theoretical grounds (Carroll 1981), either may be used to represent a general factor, and the general factor in a battery of diverse mental tests is a legitimate estimate of *g* even though it is "influenced by psychometric sampling."

In sharp contrast to the statements cited above is this later remark: "If there is doubt that the first principal factor is very similar to Spearman's *g*, . . . a finding [of high relation between factor loadings of tests on the principal factor and the size of black–white differences on those tests] would mean, at the very least, that whatever linear composite of these various tests discriminates the most among individuals *within* each population also discriminates the most *between* the means of the two populations." It must be recognized (1) that this interpretation is valid regardless of the similarity of the first principal factor (repeatedly called *g* by Jensen) to "Spearman's *g*, and (2) that Jensen seems to have forgotten by the section "Methodological desiderata" his definition of Spearman's *g* in the section "Spearman's hypothesis" (or else has decided that, contrary to his earlier declaration, Spearman's "mental energy," or "eduction of relations and correlates" is relevant, after all).

Avoidance of all such confusion would result from the total acceptance of Humphreys' definition of general intelligence (Humphreys 1971; 1984), from which it follows that a general factor score, extracted from an established intelligence test, is an acceptable estimate of a person's general intelligence. More frankly than in Jensen (1980a), Jensen now appears to accept a definition of *g* that is consistent with that proposed by Humphreys, although, as noted above, there remain some signs of conceptual slippage.

When considering the average disadvantage of U.S. black students in measures of intellectual performance, it is important to attend to the age of the students and to the year in which the data have been collected. As shown by Jones (1984), the average scores of the nation's black students on aptitude and achievement tests have steadily risen, relative to average scores for white students, over the past 15 years. Also, black–white differences have tended to be smaller for younger than for older children.

The final section of Jensen's paper reviews evidence that speed of mental processing, assessed by measuring reaction time for information-processing tasks, is consistently related to psychometric *g* and that the strength of the relation is a function of task complexity.

The evidence presented in Figure 10 is based upon 50 black students and 56 white students. It would have been useful here to find the relation between latency of processing tasks and *g* separately for the two samples, to determine the extent to which the relation is due to average black–white differences and the extent to which it is due to individual differences within each population. [Figure 10 is troublesome for other reasons as well. The equation shown for $\hat{Y}$ is actually for $-\hat{Y}$. Also, the regression line shown in Figure 10 (as well as its equation) is not the linear regression line for the data that are there displayed. A linear regression fit to those data yields a line of appreciably greater slope and lower intercept: $-\hat{Y} = -.071 + .00029X$. This regression line is far closer than the line drawn in Figure 10 to the points both at the lower left and at the upper right.]*

The data showing relations between subgroup mean latency differences and the mean latencies of processing tasks, Figures 11 and 12, are subject to a very different interpretation than that offered by Jensen. The standard deviations of latencies for the processing tasks are undoubtedly directly related to the means. Thus, Y should be standardized by the within-group standard deviations of latencies of the processing tasks. A comparable strategy was appropriately used for data presented in Figures 2, 4, 5, and 6, but is inexplicably omitted here (where heterogeneity of variance is likely to be far more severe). As presented, Figures 11 and 12 suggest that between-group mean differences are a function of within-group variability, not a surprising result.

The apparent relation of reaction time in complex processing tasks to intelligence as assessed psychometrically clearly does warrant further study. An important challenge is to try to separate the possible effects of attentiveness and motivation, which are likely to influence both response latency and test score, from the effects to be expected if the speed or efficiency of various cognitive processes is "linked to the neural substrate of mental activity."

***Editorial note:** *Please note that the author, A. R. Jensen, has been allowed to correct these technical errors in the published version of Figure 10 (q.v.), but in keeping with BBS policy that the published draft cannot diverge substantively from the draft seen by the commentators, a record of this commentator's vigilant observations on the errors in the original is here retained.*

## The nature of psychometric *g*

Paul Kline

*Department of Psychology, University of Exeter, Exeter EX4 4QG, Devon, England*

There are two issues in Jensen's target article: the investigation of the claim that the major variable differentiating the intellectual ability of blacks and whites is *g* and the investigation of the cognitive processes underlying psychometric *g*, as Jensen calls it (although the processes are related to the black–white differences). I shall deal separately with each point.

The hypothesis that Spearman's *g* was at the heart of the black–white difference in ability was tested in 11 studies by correlating the magnitudes of these differences with their *g* loadings. It is difficult to impugn the logic of the procedure: Positive correlations support the claim; anything else refutes it. In fact, the hypothesis was entirely and strongly supported.

Were there any statistical or methodological artifacts that could equally account for or, in part, contribute to these positive findings? It could be that the general factor obtained in these studies is not *g*. This is an arguable case, since there is no evidence that simple structure was obtained, and in fact *g*, in adequate rotations, usually splits into two, fluid and crystallised ability, and I should expect Spearman's argument to refer to the former. However, any failure to locate *g* accurately in factor space would have rendered the correlations smaller and thus worked *against* the hypothesis.

Since Jensen shows that other factors are far less correlated with the black–white difference and that differences between some other groups are quite unrelated to *g*, it is not possible to argue that the method per se produces positive correlations with any variables in any groups.

In brief, it seems clear that Jensen has put the issue beyond doubt and that the major determinant of black–white differences in intellectual ability is indeed *g*. Actually, if studies could have been found where tests loading on a variety of other factors had been used, thus obtaining greater variance of *g* loadings, the point would have appeared even more strongly supported, and the implication of other factors would have been clarified.

The second issue concerns the nature of psychometric *g*. Here Jensen cites research using ECTs (elementary cognitive tasks) and measures of *g*, concluding that the basis of *g* is speed of mental processing. However, the correlation of this factor with IQ was only .46. Even when this was corrected for attenuation there was less than 50% of common variance between the measures. This seems an insufficient basis for claiming that *g* reflects the speed or efficiency with which a number of elementary cognitive processes can be executed.

This aspect of the target article is less satisfactory and convincing than the first for a number of reasons. First, the ECTs used are a truly tiny sample (and a highly homogeneous one) of all the ECTs that could be employed. Carroll (1980) (who, surprisingly, was not cited), has described and classified ECTs and their relation to ability factors, and it is clear that from such a limited sample of ECTs generalisations about the nature of *g* are difficult.

Second, the account ignores the componential analysis of abilities (Sternberg 1977). The parameters revealed by componential analysis certainly indicate that processing speed is important in the solution of analogies, which are of course highly loaded on *g*. But consider the analogy that an individual cannot solve. He cannot see the relationship however long he tries; or, perhaps more pertinently, after a time of contemplation he sees the analogy. Inference, therefore, which is indeed important in *g*, seems quite separate from speed of processing.

This same argument also tends to weaken the claim that short-term memory is the essence of *g*. Clearly, if the capacity of short-term memory is exceeded, processing must break down; thus speed of processing is important. However, that inference and mapping are recognised components indicates that short-term memory is only one aspect of the cognitive processing salient to *g*.

In summary, Jensen's points are well taken, but more research is needed on the cognitive processes underlying *g*. To argue that processing speed is all is indeed too black-and-white.

## Comparative studies of animal intelligence: Is Spearman's *g* really Hull's *D*?

Euan M. Macphail

*Department of Psychology, University of York, Heslington, York Y01 5DD, England*

Jensen's target article contains three propositions that raise questions for comparative psychologists with an interest in the evolution of intelligence. The first is that there is wide variation among humans in intelligence; the second, that a major component of all human problem-solving performance is a global general intelligence factor called *g*; the third, that variations in *g* reflect variations in mental speed.

The apparent ease with which differences between humans can be demonstrated in intelligence tests contrasts with the difficulty encountered by comparative psychologists in demonstrating intellectual differences among nonhumans. My survey of the literature (Macphail 1982), for example, concluded that there currently exists no demonstration of a between-species performance difference that can be unequivocally interpreted as reflecting either a qualitative or a quantitative difference in intellect among nonhuman vertebrate species (rather than as reflecting a difference in some contextual variable such as perception or motivation). This conclusion in turn led me to suggest (Macphail 1985, in press) that we should at present adopt the null hypothesis, namely, that there are no intellectual differences among nonhuman vertebrate species. One implication, though not a necessary consequence, of this hypothesis, is that there are no within-species differences (that is, individual differences) in intellect in nonhuman vertebrates, and indeed there is no evidence for nonhuman individual differences in general intelligence (although differences in capacity for certain specific tasks – maze learning and avoidance learning, for example – are commonly reported). Now, on the assumption that human intelligence has much in common with nonhuman intelligence, it may seem odd that individual differences are found in humans alone. One possibility is, of course, that the differences seen are not in those components of the human intellect that are shared with the nonhuman intellect, but in novel components (necessary, for example, for language acquisition) that are not found in nonhumans. Another possibility is that all the individual differences in performance are due to environmental factors (some at least of which – transfer from previous experience with problems, for example – might act in comparable ways on nonhumans). A third possibility is that the human performance differences do not in fact reflect differences in intellect; this is a possibility to which I shall return when discussing the third proposition.

At first sight, it might appear that the second proposition, that there exists a major, single factor of general intelligence, is in agreement with a further implication of the null hypothesis, since if there has been no evolution of intelligence throughout the vertebrate radiation, then that intelligence is likely to be "simple" in the sense of having relatively few independent components. But the human evidence relies upon individual differences in performance – precisely the opposite of the nonhuman evidence leading to the null hypothesis. There is, in fact, no direct support for the proposal that nonhumans solving a variety of different problems use the same intellectual mecha-

nism (or mechanisms) for the solution of those problems. There is, then, no compelling reason to suppose that the two approaches have identified a common general intelligence and so no support here for the possibility that intelligence tests measure a type of intelligence common to humans and nonhumans.

The third proposition, that *g* is in effect mental speed, is potentially of the most theoretical interest, since the generation of novel attempts to test the null hypothesis relies ultimately on hypotheses concerning the ways in which (nonhuman) intellects might differ. Unfortunately, mental speed as the basis of differences in intelligence generates no proposals for potential qualitative differences in intellect and does not seem sufficiently specific to encourage formulation of new tests applicable to nonhumans (the notion that nonhuman species might differ in memory is, of course, already the subject of active investigation; e.g., Sherry 1984). Moreover, for the comparative psychologist, accustomed to be wary of contextual influences on cognitive performance, the very pervasiveness of *g*, its apparent involvement in very simple tasks – once some form of pressure is exerted – suggests that we are looking, not at an intellectual factor, but at a motivational factor, that Spearman's account of *g* as "mental energy" may be close to the truth, and that "little *g*" may be the human version of Hull's "big *D*" – the summed effect of all motivational sources, assumed to enter into and to potentiate all learned performance. It is not, of course, hard to believe that such a motivational factor would be particularly sensitive to environmental influences, and that different groups within human societies might, as a result of such influences, display characteristically different levels of such motivational energy.

## What reaction times time

T. Nettelbeck

*Department of Psychology, University of Adelaide, Adelaide, South Australia 5001, Australia*

Jensen's extensive analysis confirms Spearman's suggestion that significant mean differences in IQ between blacks and whites in the U.S.A. reflect differences in *g*. He further summarizes a substantial body of research that suggests that composite measures of timed performance might account for perhaps 25% of variance in *g*. Moreover, he provides evidence in support of the prediction that follows from these findings, of black–white differences favouring whites for IQ as conventionally measured as well as on various speed measures. Thus his conclusion that real differences in *g* exist is plausible. However, the explanation that he advances as to how such differences arise is not convincing.

It is clear that Jensen regards *g*, and hence the black–white differences in question, as being predominantly genetically determined, a position consistent with the hereditarian model of intelligence that he has advanced over at least the past 15 years. In the present article he argues that his position is strengthened by finding differences in reaction time (RT) and similar indices of speed of performance. This argument is based on the assumption that RT measures basic perceptual and memory capacities, fundamental capacities in the sense that complex intellectual functions influenced by learning are not involved. This assumption is not correct.

Even among samples confined to the university and college students frequently employed in much RT research, it is not recognized that practice produces significant decreases in RT. This is because subjects learn to use increasingly more efficient strategies (Rabbit 1979; Salthouse & Somberg 1982). Such effects are found in tasks of the kind described by Jensen, being less pronounced for light-key tasks (refer to his Figure 8) than for digits and words as reaction stimuli, as illustrated in his Figure 9 (Teichner & Krebs 1974). Jensen has previously reported absence of practice effects for these tasks (e.g., Vernon & Jensen 1984), but this conclusion was derived from internal consistency measures made from only 26 trials in each instance, arguably an insufficient number to permit practice effects to emerge. It is also now well established that criterial factors can determine choice RT by influencing the trade-off between the speed and the accuracy of responding (Pachella 1974). Jensen's procedures have been checked in this regard, and the possibility of a speed–accuracy trade-off has been discounted (Vernon 1983), but only because of virtually zero phi correlations between errors and RT dichotomized as being above or below a median value; no error rates have actually been published for any of this work. However, Nettelbeck and Kirby (1983) measured two-, four-, six-, and eight- choice RT using equipment and procedures modelled on Jensen's (refer to his Figure 8) for 141 nonretarded and 41 mildly mentally retarded young adults. They found that RT errors with this apparatus were extremely rare, amounting to less than 0.5% in both samples. A near-zero error rate is undesirable in RT research, since the experimenter cannot be confident that subjects are responding at optimum speed, just at the point on the speed–accuracy operating function where minimum time is taken to achieve perfect accuracy. Even approaching this optimum, RT is near asymptote, and very large changes in speed are possible in exchange for barely discernible changes in accuracy.

Practice and criterial variables of the kind alluded to would not be critical to Jensen's causal explanation if one could assume that such variables would not disproportionally influence individual performance or the average outcome within different groups or populations. However, this assumption is not justified because individual and group differences in cognitive variables reflecting emotional factors or influencing attitude, persistence, and other motivational considerations capable of affecting reaction efficiency are certainly possible and have been reported (Carlson & C. M. Jensen 1982). Jensen's finding (e.g., Jensen 1980b) of strong correlations in some samples between IQ and the nondecision movement component of the reaction is contrary to his predictions but could plausibly be attributed to some third variable associated with both IQ and RT, like attention or motivation. Consider too the study of Roth (1964), who first applied Hick's variable choice RT method to the investigation of intelligence, comparing intellectually able and handicapped children. As predicted by a speed model of intelligence, the slope of the regression of RT on "bits" of information correlated significantly ($r = -0.39$) with IQ, whereas the zero intercept of the regression line did not. However, Roth (1964) also reports a significant correlation of $-.41$ between the slope and the intercept of the regression function. This outcome, which is inconsistent with a speed model of intelligence, has been confirmed by Nettelbeck and Kirby (1983). It suggests that subjects apply different criteria for responding at different levels of choice.

A recent study by Borkowski and Krause (1983) has provided evidence consistent with an environmental explanation for black–white differences in both IQ and RT. Unexpectedly, in view of Jensen's results, this study failed to find significantly slower choice RT among black eight- and nine-year-old children who scored below white children on tests of fluid and crystallized intelligence, although differences were in the predicted direction. However, significant sample differences in simple RT were well accounted for by differences in executive components of processing, particularly those reflecting general knowledge and metacognition. RT methods provide useful ways for analyzing how persons of different ages and abilities attempt to solve problems of discrimination and judgment. One cannot simply assume that the processes involved are reduced to fundamental levels of biological efficiency. Jensen's important findings emphasize the urgent need for social and educational programs aimed at counteracting black–white differences in *g*. Borkowski and Krause's results suggest a future direction.

## Intelligence and its biological subtrate

Robert C. Nichols
*Department of Educational Psychology, State University of New York, Buffalo, N.Y. 14260*

Psychologists do not agree on the underlying structure of human abilities. Two conceptualizations fit the results of factor analyses of psychometric tests equally well: (1) a large general factor of intelligence common to all abilities with a number of smaller factors specific to particular test content, and (2) several primary mental abilities specific to particular mental operations that may be more or less correlated with each other. The continuing conflict between these two approaches began when Thurstone's (1938) primary mental abilities were first contrasted with Spearman's (1927) theory of *g*. The major unresolved issue seems to be whether the observed general factor is due to individual differences in some basic biological structure or process responsible for intelligence, or whether it is the result of imprecise measurement and correlated environmental influences.

Jensen (1980a) has made the most extensive and thorough statement yet in favor of the *g* theory, and the arguments are adequately summarized in the present target article. Jensen (1973a) has also made the most extensive and thorough statement yet concerning ability differences between blacks and whites in the United States. The target article brings these two lines of research together with Spearman's hypothesis that the black–white difference is primarily a difference in the general factor. The empirical support for this hypothesis reported by Jensen seems more than adequate. In fact, the evidence is so strong and pervasive that the impressive technical sophistication of the analysis hardly seems necessary.

It is interesting to note that previous studies of the pattern of population differences in ability, cited by Jensen, have emphasized the primary-mental-abilities approach, with the provocative finding that the shape of the profile of abilities appears to be a stable characteristic of black, white, and other ethnic samples across different socioeconomic levels. Jensen has now shown that, for the black–white comparison at least, the difference in profile shape is largely accounted for by the different *g* loadings of the tests.

The interpretation of this finding depends on one's position concerning the nature of the general factor. If the general factor is intelligence it means one thing, but if the general factor is biased measurement and correlated environmental inputs to the more basic primary abilities it means another. Thus, attention is focused on Jensen's argument that *g* is intelligence.

The cumulative weight of the evidence summarized by Jensen under the heading "the nature of *g*" may appear overwhelming to some, but it has not led to general agreement in the past. Nevertheless, two lines of evidence, if true, would seem to compel acceptance of *g* as "*the* primary mental ability." These are (1) the inbreeding depression of *g* more than of other abilities, and (2) the substantial correlation of *g*, and of no other ability independent of *g*, with basic speed of mental processing. These last two critical lines of evidence are new additions to the argument that have been contributed primarily by Jensen and his associates. They deserve to be replicated and explored by others.

In particular, the relationship between psychometric tests and speed of mental processing of elementary cognitive tasks deserves more extensive investigation. The importance of this line of research goes far beyond the issue of black–white differences, since the nature of intelligence itself is at the end of this particular rainbow. A new discovery of "gold" has been reported in California: the first substantial connection between intelligence and its biological substrate.

Frankly, the reported correlations of around .70 (after legitimate corrections) between *g* and speed of mental processing seem too good to be true. A correlation half that size between psychometric test scores and some basic mental operation would be enough to get very excited about. This promising lead should be aggressively followed up to see whether it is, in fact, the significant breakthrough it appears to be.

## Empirical evidence of bias in choice reaction time experiments

Ype H. Poortinga
*Department of Psychology, Tilburg University, Tilburg, Netherlands*

For a meaningful intergroup comparison it is essential that in the groups concerned the score variable form an "equivalent" or "unbiased" scale of the psychological function or trait to which the scores are being generalized. Similarity of correlation matrices or factor structures across groups can be considered a necessary condition for equivalence, but it is not a sufficient condition. For example, the effect of a lack of equivalence cannot be detected by means of correlational analysis if it can be expressed as a linear function of the observed score variable. Also, the absence of a significant stimulus by group interaction in an analysis-of-variance design is not a sufficient condition. A bias effect that approximates a constant across the subjects in a group will not show up in an interaction but will show up in the main effect for groups. In general, it is difficult to make plausible that intergroup differences are not caused by metric inequivalence of tests, epecially when inferences are made about broad domains of behavior (Poortinga 1983; Van de Vijver & Poortinga 1982).

Cross-cultural studies have shown time and again that the content of items is an important determinant of performance level on cognitive tests. Cross-cultural differences on traditional intelligence tests are influenced by the familiarity of the subjects with the kinds of operations required and the opportunity to learn certain items of knowledge and certain skills (e.g., Cole, Gay & Glick 1968; Luria 1976; Ombrédane, Robaye & Plumail 1956; Serpell 1979). One strategy to make less biased estimates of intergroup differences is to reduce the role of culture-specific experiences. Jensen is following this strategy when he refers in his target article to intergroup differences in choice reaction time (CRT) and similar tasks. Jensen's arguments require (1) that *g* be identifiable with information-processing capacity as measured by CRT tasks and (2) that these tasks not show biased results (in terms of information-processing capacity) across groups. My concern here is with the latter requirement.

According to Jensen the most important parameter for individual differences is the rate of increase in RT with increasing complexity of the stimuli or with an increasing number of stimuli in the task. I shall call this parameter *a*. Several factors have been identified in RT studies that affect *a*, such as speed–accuracy trade-offs, stimulus–response compatibility, stimuli's discriminability and training (Fitts & Posner 1967; Welford 1980).

In 1971, I reported a study on auditory and visual information transmission with two groups of South African university students. There were 40 African subjects and 40 subjects of European descent, each group consisting of an equal number of men and women (Poortinga 1971). Part of the study consisted of a CRT experiment with three conditions: a four-choice auditory task, a four-choice visual task, and an eight-choice task during which the two sets of stimuli were presented within the same series. Subjects used the same four push buttons during all three conditions, keeping a finger on each button all the time. There were also simple RT tasks for clicks and flashes, but these were administered apart from the CRT experiment.

The main results are presented in Figure 1. There is a clear difference across groups in the increase in response time from the simple RT to the four-choice RT, in both the auditory and

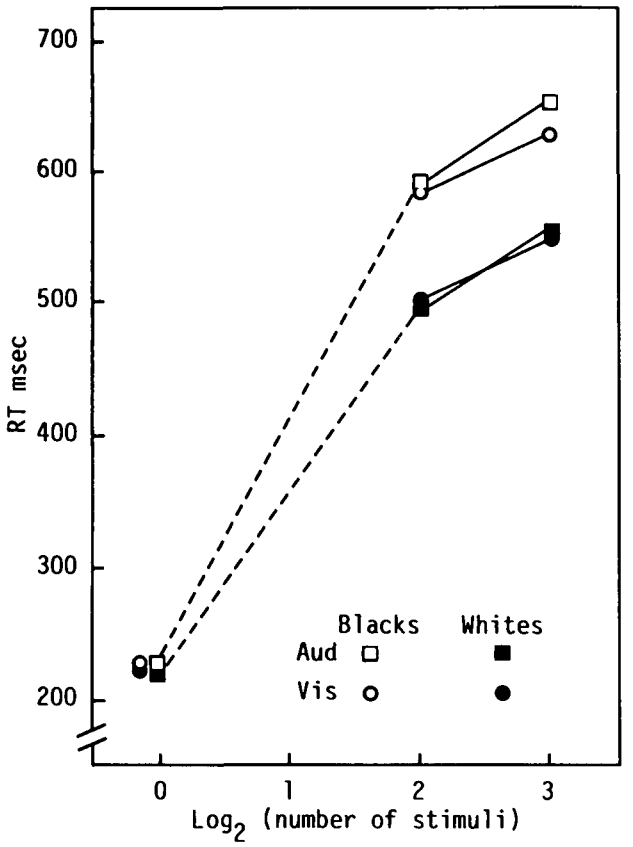


Figure 1 (Poortinga). Results on RT and CRT tasks performed by black and white South African students. Each entry is based on the mean of the distribution of median RTs in a group. (After Poortinga 1971.)

the visual task. Jensen (1980a, p. 705) has referred to this result and interpreted it as providing support for his ideas. However, this interpretation can be challenged.

First, the increase in CRT from the four-choice to the eight-choice tasks is very similar across the groups, which means that the values on the *a* parameter in the CRT experiment are about equal across groups. Therefore, this estimate is (in Jensen's framework) incompatible with the result Jensen has emphasized. Second, and more important, the differences in the *a* parameter can be at least partly explained in terms of stimulus specific factors. For both the auditory and the visual CRT tasks a significant interaction effect between stimuli and culture was observed (Poortinga 1971, pp. 49, 54). This result indicates that the relative difficulty of the stimuli within a task was not identical for the subjects in both groups. What the effect was on the overall level of performance is difficult to decide. Since the response time for a given stimulus cannot be taken as independent of the other stimuli in a task, an overall effect on performance certainly cannot be excluded. Moreover, in the study discussed here it was observed that with training the intergroup differences between stimuli became smaller (Poortinga 1971, p. 60).

As a point of interest it can be mentioned that on two information transmission tasks in which speed was not a factor, namely a loudness and a brightness-judgment task, only small differences between the same two groups of subjects were found (Poortinga 1971, p. 41).

In conclusion, any suggestion that CRT tasks are highly culture-free cannot be maintained as long as the effects of psychometric inequivalence can be clearly demonstrated. It appears that this argument cannot be invalidated with tasks with high stimulus–response compatibility, such as these used by Jensen (c.f., e.g., Jensen 1982b). Such tasks will presumably show a low value for the *a* parameter in any group and consequently only (numerically) small intergroup differences on this parameter. The question still remains whether these small differences should be associated with "*g*" or "intelligence" or other factors, which, in terms of intelligence, make the parameters nonequivalent.

## Chronometric measures of *g*

Michael I. Posner

*Department of Psychology, University of Oregon, Eugene, Ore. 97403*

I would like to confine my commentary on Jensen's article to the experiments involving reaction-time tasks. It seems to me there is a mistaken impression that these tests do a better job of eliminating differences due to motivation and past education than they actually do. Moreover, insofar as the results relate to the speculation that black–white differences in intelligence test scores are due to basic information-processing capabilities, they fail to provide any evidence for it. Indeed an analysis of Figures 11 and 12 might lead to exactly the opposite conclusion.

For example, it may seem that speeded matching of identical words is so simple that little improvement with specific learning would be possible. It has been well established, however, that such judgments are heavily influenced by factors such as orthographic regularity (Carr, Posner, Pollatsek & Snyder 1979) that are critically dependent on the amount of specific reading skill a person has acquired. Chi (1976) has provided evidence that much or perhaps all of the improvement in the memory span that occurs with age can be attributed to improvement in the speed of processing digits due to increased familiarity. Studies of the reaction times of aged persons show how critically the differences in speed of response occurring with age are related both to practice and to motivation to avoid fast times in order to reduce errors (Rabbitt 1981).

Most users of these methods are aware of such problems and use some version of the subtractive method to reduce general effects of motivation and to isolate elementary operations. Thus one could plot the slope of the function relating reaction time to number of alternatives, or the time differences between physical and name matches. Jensen neglects to do so with these data, thus increasing the degree to which performance might rest on general factors of motivation to perform rapidly. In addition, the use of overall task scores makes it difficult to argue that any task represents a relatively isolated operation. This reduces the likelihood that these experiments will aid in the development of a theory of intelligence.

In addition to these general problems with the chronometric studies there is also reason to doubt the degree to which the specific data obtained support the assertions of a difference in speed of general information processing between blacks and whites. Figure 12 compares students in vocational colleges and universities. Presumably these students differ in many ways, including measured intelligence and degree of education. The eight subtests show substantial advantages for the university students. Excluding task 1 these range from 100 to 400 msec and are ordered exactly as one would expect based on the difficulty of the test as measured by reaction time.

An examination of Figure 11 (black–white differences) yields a quite different result. The correlation between task difficulty and group differences drops from .98 to .74. The slope drops markedly so that the difference between groups is at most 100 rather than 400 msec. Moreover, the task ordering appears quite different from what would be expected from the idea that the more information processing involved the greater would be the advantage of whites. Instead it appears that there is little or no difference between whites and blacks except where the test involves reading words. It is true that test 5 requires the subject

merely to determine whether the pair is physically the same, but, as suggested above, this judgment is influenced by the subject's visual knowledge of words. Of course the data of Figure 11 do not provide any definitive evidence that the racial differences might best be related to time spent in reading, but this pattern of data is at least as consistent with the differential reading skill hypothesis as with the idea that they represent some basic fundamental information-processing speed.

Note the clear difference in the degree of sensationalism involved in these two ways of expressing the result. To conclude that black students have less experience in reading or have spent less time reading than white students is of some interest, but most of us would not be surprised by it. On the other hand, to conclude that blacks have slower systems underlying information processing in all forms entails a highly emotion-ridden and controversial position. These chronometric data are more consistent with the less controversial form of the hypothesis.

Although reading skill is important in our society and might turn out to have a lot to do with the correlation among many intelligence subscales (e.g., *g*), it is not reasonable to suppose that it is a general property of the brain or reflects general ability to process information. The data on acquired dyslexias (Coltheart 1981) argue clearly that one can interrupt a specific component of the reading process (e.g., developing a phonological description) while leaving the rest intact. The thrust of current work in cognitive psychology and cognitive neuropsychology is to associate the performance of different elementary operations with quite separate neural systems (Posner, Pea & Volpe 1982). Most of this work suggests that instead of a single integrated factor, operations within different cognitive systems draw upon separate resources.

In short, chronometric experiments do not necessarily provide evidence of basic mental operations free of past experience and current motivation. Used carefully they might provide a theoretical basis for an understanding of intelligence, but the studies in this paper are not constructed to do so. These results do show a relationship between overall task difficulty as measured by reaction time and differences between vocational and university students. However, the black–white differences are much smaller and more variable and appear to be more closely related to differences in reading skill.

## Oh *g* Dr. Jensen! or, *g*-ing up cognitive psychology?

P. M. A. Rabbitt

*Age and Cognitive Performance Centre, University of Manchester, Manchester M13 9PL, England*

Any hypothesis of racial differences invoking a "single common factor" is obviously convenient for genetic theories of intelligence, but Jensen has to face the difficulty that "not every *g* is an equally good *g*." His paper undertakes to habilitate *g* as not merely "a theoretically empty artifact of factor analysis," or an aphoristic "property of the mind" (Wechsler quoted by Jensen) but a construct in cognitive psychological models of perceptual motor processes and (perhaps even more convenient for genetic theories) as an index of neurophysiological competence. Can we agree?

Jensen's grasp of cognitive process models is variable. He takes the position that if we cannot do some things quickly we may not be able to do them at all. This is based on a 1960s idea of "immediate memory" as a passive store, in which necessary "information" will rapidly decay if slow concurrent decisions preclude rehearsal. Thus, to do many things well (i.e., to be clever) it may be necessary to be quick. A less eclectic theorist who took the term "working memory" at the value given it by its inventors Baddeley and Hitch (1974) would attribute differences in span to relative speed of rehearsal (as in an "articulatory loop"; c.f. Hulme, 1984) and so would find Jensen's observation that (low-*g*?) blacks perform as well as, or better than, (high-*g*) whites at memory tasks such as rote learning or digit memory span uncongenial for a proposed *simple* equivalence of information-processing speed and *g*. However, Jensen distracts himself by plotting forward and backward memory span against socioeconomic status (0 to 9) separately for blacks and whites. For both populations, both spans increase with group socioeconomic status. There is a white advantage only for backward span. Jensen concludes that this is because backward span is twice as heavily loaded for *g* and that *this* is because backward span requires retention of forward-presented digits in working memory while they are (more or less rapidly) transformed for backward recall. The less *g*, the slower the transforms, and the greater the probability of loss and error.

The implication is that two populations carefully matched on a low-*g* task nevertheless differ on a high-*g* task. This "matching" is dubious. All inhabitants of a telephone/computerised/numerical-coding culture get daily, highly motivated practice at forward span, but much less at reordering remembered sequences. Given that Chase & Ericsson (1981) has shown that practice can improve forward span above 200 to no known upper limit, daily practice must contract differences between individuals to fit cultural demands. Individual differences will appear only in unpractised tasks such as backward span. It suits this hypothesis that differences *between socioeconomic groups* are also much smaller for forward than for backward span. Interpretation of black–white differences on backward digit span thus hangs *solely* on our assumptions as to what it actually means to "match" ten pairs of groups of blacks and whites, rank-ordered for "socioeconomic status."

Jensen's main evidence is that on a battery of six different tests of decision speed yielding eight separate performance indices (Vernon 1983) decision speed is inversely related to *g* and that this relationship becomes stronger as tasks become more complex and slower. There are difficulties:

1. Jensen and Vernon adapted all the reaction-time (RT) tasks they borrowed from the cognitive literature by requiring volunteers to keep one finger on a "home" button, moving it (unstated) centimeters in Task 1, cf. Figure 8; or 10 cm in all other tasks, cf. Figure 9) to touch the appropriate one of two to eight response keys as soon as possible after a signal was presented. Only times between signal onset and liftoff from the home key are analysed here. This seems silly because:

(a) Especially if reinforced for speed by feedback of their liftoff times, volunteers might make a fast liftoff on display onset followed by a "hover" to evaluate it. Moreover, Rabbitt and Rodgers (1965) have shown that a display can be evaluated during a reach. Only determined compliance and long practice would eliminate this tendency, which would show up in difficult tasks (which are the only ones hinting at correlations with black–white differences and, inferentially, *g*).

(b) Especially in the more complex tasks, this questionable methodology complicates experimental instructions and introduces gratuitous elements of self-monitoring and compliance, which, when pathetically little practice is given, must be the most powerful factors affecting individual performance. For example, people with high Armed Services Vocational Aptitude Battery (ASVAB) scores, who might be familiar with abstract procedures using electromechanical equipment, would have a great advantage in task adaptation (see below).

(c) Volunteers would correct many, if not all, errors by unscorable mid-flight reach adjustments. Such "corrected errors" would be pooled to yield unusually fast "liftoff" times for especially impulsive volunteers.

2. It is therefore unsurprising that Jensen reports no data on errors. But this makes it impossible to know whether groups differed in willingness to trade accuracy for speed.

3. Volunteers received remarkably little practice on any task. For example, there were only 26 trials on each of the difficult

SD2 and SA2 tasks, on which the claimed correlations chiefly depend. The *maximum* for any task was 84 trials on Memory Scan (i.e., 12 trials per span length!). This makes the data uninterpretable because:

(a) When so few observations are made, differences in mean RTs are uninformative because they mainly reflect differences in the variance and kurtosis of RT distributions. Often volunteers' *fastest* responses differ little, or not at all, between groups or conditions so that differences in means are determined entirely by isolated slow responses.

(b) In an unfamiliar task the first one to ten responses a person makes on an unfamiliar task may be two to six times slower than those produced when verbal instructions have been worked through in terms of their physical implementation. The more difficult the task and, no doubt, the less sophisticated the volunteer, the longer this settling-down period will take. This raises the fundamental question of what the differences between Vernon's tasks really measured, differences in the times taken to come to terms with quite complex instructions or differences in information-processing speed?

(c) This in turn raises the more general question of precisely what claim Jensen intends. Even very modest amounts of practice reduce mean RTs by 100 to 300%. Improvement with practice continues after periods of 25 days or more. A finding that differences in ASVAB scores predict differences in times taken to learn unfamiliar tasks in a strange social context is not very informative. To test a claim that differences in *g* reflect functional, even perhaps neurophysiological, differences, we must compare groups at asymptotic performance. Neither this nor any other study Jensen quotes separates the trivial from the interesting possibility.

Among many lapses of logic and questionable assumptions, the following are notable because they appear in other studies than those cited here:

(a) Where the outcome favours his hypothesis, Jensen punctiliously adjusts correlations to take account of gross differences in ranges of scores (e.g., in the penultimate paragraph of the section "Information-processing capacities and psychometric *g*"). He makes no adjustments for what must have been gross increases in variance between the difficult tasks (e.g., SD2 and SA2 over the easiest tasks 1, 2, 3, 4, and 6; cf. Figure 10). His argument depends entirely on this putative difference.

(b) Though "reaction time" is *measured* in constant units of milliseconds it is not *functionally* an equal interval scale; that is, a shift between mean RTs from 180 to 280 msec is not *functionally* equivalent to one from 1000 to 1100 msec. Plots of RT against condition difficulty sometimes appear linear over a brief range (e.g., Sternberg 1969) but more often accelerate or decelerate to an asymptote. Interpretation can be made only in the light of functional models after careful task analysis.

With these points in mind it seems supererogatory to go on to inspect the actual data; however:

1. Why are intercorrelations between the tasks, and their possible variance across groups, not given? It seems likely that the reading-based tasks (SD2 and SA2) would correlate rather poorly with the others.

2. Why does Jensen find it reassuring that the ASVAB "coding speed test" correlates only weakly with his battery? In my own experience this test predicts performance well on a variety of visual search tests and other measures of information-processing speed. To my mind the absence of a correlation validates objections to Jensen's methodology.

3. Correlations with the ASVAB *g* factor are unimpressive for tasks 1, 2, 3, 4, and 6. Do we count this as a failure of replication of correlations of 0.4 and above, for task 1, cited in Jensen (1981; 1982d)?

4. The overall correlations evidently depend substantially on tasks (SA2; DT3 words) similar to those which Hunt and associates have shown to be related to verbal ability. Why are excellent experimental series such as Hunt, Lunneborg & Lewis (1975); which are unfavourable to the pure *g* hypothesis, not cited?

5. Why is Jensen excited by dual task correlations with ASVAB *g*, since these are no better than those obtained when component tasks are administered in isolation ? (cf. the near parity of 3 and 4, of 5 and 6, and of 7 and 8 in Figure 10.) This *failure* to find increased correlations between task performance and *g* scores in complex tasks, involving overall superordinate control of processing, is very unfavourable to Jensen's argument and to Vernon's methodology. It also strongly hints that the relatively high correlations between "ASVAB *g*" and SD2 and SA2 scores, whether they appear as tasks in isolation or components in dual tasks, reflect their verbal content rather than any intrinsic information-processing difficulties they entail.

This is not a convincing paper. Excellent reviews by Cooper and Regan (1982), Hunt (1978), and R. Sternberg (1982) show that mapping psychometric models and concepts onto process models developed by cognitive psychologists may now be one of the most important goals for cognitive science.

## Differential K theory and group differences in intelligence

J. Philippe Rushton

*Department of Psychology, University of Western Ontario, London, Ontario, Canada N6A 5C2*

The difference between blacks and whites in the United States on measures of intelligence has remained at approximately one standard deviation for the last 70 years (Loehlin, Lindzey & Spuhler 1975). Jensen's detailed and scholarly treatment is important because it convincingly addresses the nature of this difference. This commentary builds on his discussion of group differences to include Asians, countries beyond the United States, and traits in addition to intelligence. At the conclusion, "differential K theory" is described to organize the observations within an evolutionary framework.

***Intelligence.*** Some Asian people score higher on tests of intelligence than Europeans. Despite peasant background and initial discrimination, on average the Chinese and Japanese in Canada and the United States have reached higher educational and occupational levels than Euro-Americans, and they score higher on tests of intelligence (Vernon 1982). Other studies document the higher intelligence of the Japanese in Japan (Lynn 1982, but see Flynn 1984; Misawa, Motegi, Fujita & Hattori 1984). People of African descent, however, score lower than Europeans on measures of intelligence elsewhere in the world, including Britain (Scarr, Caparulo, Ferdman, Tower & Caplan 1983), and such postcolonial African countries as Nigeria, Tanzania, and Uganda (Lynn 1978). If the cultural attainments of Asians, Europeans, and Africans on their home continents are examined (e.g. by dating such inventions as written language, numbering systems, calendars, astronomical systems, codified rules of law, domestication of plants and animals, and metal technology), the rank ordering remains the same (Baker 1974).

***Activity level.*** Newborn Chinese-Americans, on average, are quieter and more readily soothed than Euro-Americans who, in turn, are less active than Afro-Americans (Freedman 1979). One measure involves pressing the baby's nose with a cloth, forcing it to breathe with its mouth. Whereas the average Chinese baby appears to accept this, the average Euro- or Afro-American baby fights it immediately. Subsequent studies have replicated these findings in other countries with quite different measures and samples. The Navajo Indians of the southwestern United States, for example, stoically spend much of their first six months of life wrapped to a cradleboard. Attempts to get Euro-American infants to accept the cradleboard have met with little success (Freedman 1979). The Navajo are like the Chinese in being classified as belonging to the Mongoloid population.

***Behavioral restraint.*** A large number of studies have tested the personality of the Chinese and Japanese both in their homelands and in North America (Vernon 1982). On questionnaires, Asians are, on the average, more introverted and anxious and less dominant and aggressive than Europeans. These differences are manifest in play behavior, with Asian children being quieter, more cautious, and less competitive and aggressive than Euro-Americans. Eskimos, who are also Mongoloid, are likewise behaviorally restrained (LeVine 1975). African-descended people, on the other hand, tend toward the extraverted end of the continuum. Individual differences in anxiety, behavioral restraint, and extraversion have been linked to the inhibitory system of the brain (Gray 1982). [See also Zuckerman: "Sensation Seeking" *BBS* 7(3) 1984.]

***Developmental precocity.*** In the United States, blacks have a shorter gestation period than whites. By week 39, 51% of black children have been born, while the figure for whites is 33%; by week 40, the figures are 70% and 55%, respectively (Niswander & Gordon 1972). This precocity continues throughout life. In terms of physical coordination, Freedman (1979) found that, unlike Europeans and Asians, many African as well as Afro-American newborns can hold their heads erect. Concomitant differences are found in skeletal maturity, as measured by growth of ossification centers throughout the first years of life (Eveleth & Tanner 1976). Afro-American children also walk at an average age of 11 months, compared with 12 months in Euro-Americans, and 13 months in American Indians (Freedman 1979). Afro-Americans are also more precocious sexually, as indexed by age at menarche (Malina 1979), first sexual experience (Weinrich 1977), and first pregnancy (Malina 1979).

***Differential K theory.*** In the discussion above, Europeans fell midway between Asians and Africans. The ordering raises interesting theoretical questions, especially since there is evidence for the heritability of the traits discussed, including intelligence (Bouchard & McGue 1981), activity level (Willerman 1973), behavioral restraint (Floderus-Myrhed, Pedersen & Rasmuson 1980), rate of growth (Wilson 1983), age at menarche (Bouchard 1982) and age of first sexual experience (Martin, Eaves & Eysenck 1977). Differential K theory has been proposed to help order these and other biosocial differences found between people (Rushton 1984a; b; 1985).

It is postulated, on the basis of concepts from evolutionary biology, that the degree to which an individual engages in a "K" reproductive strategy underlies multifarious characteristics related to life history, social behavior, and physiological functioning. K refers to one end of a continuum of reproductive strategies organisms can adopt, characterized by the production of few offspring with a large investment of energy in each. (K is a symbol from population biology, standing for the carrying capacity of the environment, or the maximum population a species can maintain under certain fixed conditions.) At the opposite extreme is the r strategy in which organisms produce numerous offspring, but invest little energy in any one. (r is also a symbol from population biology and stands for the maximal intrinsic reproductive rate, or the natural rate of increase in a population temporarily freed from resource limitations.) Oysters, producing 500 million eggs a year, exemplify the r strategy, while the great apes, producing only one infant every five or six years, exemplify the K strategy. Across-species comparisons demonstrate that a variety of life history features correlate with these reproductive strategies, including litter size, birth spacing, parental care, infant mortality, developmental precocity, life span, intelligence, social organization, and altruism (Wilson 1975).

As a species, humans are at the K end of the continuum. Some people, however, are postulated to be more K than others (Rushton 1985). The more K one is, the more one is likely to be from a smaller-sized family, with a greater spacing of births, a lower incidence of dizygotic twinning, and more intensive parental care. Moreover, one will tend to be more intelligent, altruistic, law abiding, behaviorally restrained, maturationally delayed, lower in sex drive, and longer lived. Thus diverse organismic characteristics, not apparently otherwise related, are presumed to covary along the K dimension. With respect to group differences, Asians are hypothesized to be more K than Europeans, who, in turn, are hypothesized to be more K than Africans. This ordering accords well with data on multiple birthing, which can be taken as an index of litter size. For example, the dizygotic twinning rate per 1,000 births among Asians is 4; among Europeans, 8; and among Africans, 16 (Bulmer 1970). Similarly, a comparison of the incidence of triplets and quadruplets shows a higher frequency among Africans than Europeans (MacGillivray, Nylander & Corney 1975). A parallel ranking in longevity has been found (Bengtson, Kasschau & Ragan 1977). Numerous other indices of K correlate both between and within populations (Jensen 1984d; Rushton 1985). The nature of black–white differences in *g* may belong in a broader evolutionary context than has been considered to date.

## Neural adaptability: A biological determinant of *g* factor intelligence

Edward W. P. Schafer

*Brain-Behavior Research Center, University of California, San Francisco, Sonoma Developmental Center, Eldridge, Calif. 95431*

This commentary addresses Jensen's statement that "little, if anything is, as yet, known about the physiological and biochemical substrate of *g*."

Our studies of evoked cortical potentials have identified significant brain electrical activity differences that could account for human variability in *g* factor intelligence (Schafer & Marcus 1973; Schafer 1982; Schafer 1984). The working hypothesis for these studies has been that individual differences in the cognitive modulation of evoked potential amplitude will relate to individual differences in behavioral intelligence.

In the 1982 study, auditory evoked potentials (EPs) were obtained from 109 normal and 52 mentally retarded adults under three stimulation conditions (periodic, self, and random) designed to manipulate temporal expectancy. The normal adults showed a strong temporal expectancy effect on their EPs, giving smaller than average EPs to expected inputs and larger than average brain responses to unexpected stimuli. In contrast, the retarded adults failed to show a temporal expectancy effect on their EPs, indicating a deficit in cognitive neural adaptability. A measure of neural adaptability derived from EP amplitude ratios correlated .66 with WAIS IQ scores obtained on 74 normal adults, indicating a definite association between neural adaptability and behavioral intelligence. This correlation rose to .82 when corrected for the restricted range of IQ (98 to 135) in the sample. People who gave larger than average EPs to unexpected inputs and smaller than average EPs to stimuli whose timing they knew tended to have higher IQs. Results suggested that the brain that efficiently inhibits its response to insignificant inputs and that orients vigorously to unexpected, potentially dangerous stimuli is also the brain that manifests high behavioral intelligence. Neural adaptability as indexed by the temporal expectancy effect on evoked cortical potentials appeared to provide a biological determinant of *g* factor psychometric intelligence.

If the EP temporal expectancy index is a good measure of *g* factor intelligence, then WAIS subtests having high *g* factor

loadings should also show high correlations with the EP measure, while subtests having low *g* factor loadings should show low correlations with the EP temporal expectancy index. A rank-order correlation of .71 (p .01) was obtained between the WAIS subtest *g* factor loadings and the correlations of the same subtests with the EP temporal expectancy index. The WAIS subtest *g* factor loadings in this sample showed a coefficient of congruence of .97 with the loadings of the same subtests with the first principal factor from the standardization sample reported by Matarazzo (1972), indicating that this factor was indeed Spearman's *g*. This degree of correspondence indicates that whatever is measured by the subtests of the WAIS, presumably *g* factor intelligence, is also measured by the EP temporal expectancy index.

Eysenck and Barrett (1985) have suggested that the theory of cognitive neural adaptability as a basis for behavioral intelligence could be more elegantly demonstrated using an EP habituation paradigm. Consequently, a recent study (Schafer 1984) tested the hypothesis that the magnitude of EP habituation could relate to individual differences in psychometric intelligence.

In this study, 19 male and 33 female subjects listened to 50 moderately loud click stimuli presented with a fixed 2-second interstimulus interval. Vertex EPs were averaged off-line for the first and second blocks of 25 stimuli. The percentage difference between the two averages in the amplitude of the N1-P2-N2 excursion served as the measure of EP habituation.

The EP habituation index correlated .59 (p .001), with WAIS full scale IQ indicating that the greater the EP habituation the higher the behavioral intelligence. When corrected for the restricted range of IQ scores in the sample (98–142) the correlation rose to .73. By using the 15% EP habituation measures as a cutoff point, 75% of the subjects were correctly identified as having either superior (120+) or average IQs. Combining the habituation index and an index of EP temporal expectancy as predictors resulted in a higher correlation (.64) with the criterion variable WAIS IQ than either predictor taken separately. This multiple correlation rose to .80 when corrected for the restricted range of IQ, a correlation higher than that of one IQ test with another.

The results indicate an association between the degree of EP habituation and behavioral intelligence, suggesting that the brain that efficiently inhibits its response to repetitive, insignificant inputs is also the brain that shows high behavioral intelligence.

Again, if the EP habituation index is a good measure of *g* factor intelligence, then WAIS subtest *g* factor loadings should correspond with the subtests' correlations with the EP measure. A rank-order correlation of .91 was obtained between the WAIS subtest *g* factor loadings and the correlations of the same subtests with the EP habituation index. The WAIS subtest *g* factor loadings in this sample showed a coefficient of congruence of .98 with the loadings of the same subtests with the first principal factor from the standardization sample reported by Matarazzo (1972), indicating that this factor was Spearman's *g*. This notable degree of correspondence indicates that whatever is measured by the subtests of the WAIS, presumably *g* factor intelligence, is also measured by the EP habituation index.

Neural adaptability as indexed by the habituation of evoked cortical potentials and the temporal expectancy effect on these potentials provides a biological determinant of *g* factor psychometric intelligence. Given these observations and those of other workers studying EP correlates of IQ (Hendrickson & Hendrickson 1980), we can agree with Jensen that Spearman's *g* is not merely an empty mathematical artifact of factor analysis but rather a construct possessing biological significance. By identifying correlates of *g* factor intelligence outside the psychometric realm the evoked potential studies may help to elucidate the essential nature of *g* and hence of human intelligence.

## On artificial intelligence

Peter H. Schönemann

*Ludwig Maximilian Universität, München, Federal Republic of Germany and Department of Psychological Sciences, Purdue University, West Lafayette, Ind. 47907*

1. In his target article, Professor Jensen has adduced impressive empirical evidence in support of "Spearman's hypothesis" that the elements of the eigenvector associated with the largest latent root of a correlation matrix of "intelligence tests" correlate positively with the mean white–black differences on these tests. He believes that this finding sheds further light on the nature of the black–white difference on various psychometric tests (the title of the target article) and the "nature of g"; he has discussed "the practical implications of *g* and Spearman's hypothesis for employment, productivity, and the nation's economic welfare . . . in more detail elsewhere."

It will be shown in this commentary that the predicted correlation has nothing to do with any of these things because it is a psychometric artifact that arises with any data as long as the covariance matrices are equal and the mean vectors are sufficiently different.

2. Let $e' = 1,1, \ldots, 1)$ be a row vector of N ones (so that $e'e = N$) and consider two Nxp ($N > p > 1$) score matrices X, Y, drawn from two populations that differ only in the mean vectors, not in the covariance matrices. If the means in the 2Nxp pooled score matrix $Z = (X';Y')'$ are set to sero, these two score matrices X, Y, can be written

$$X = U + ed', \; Y = V - ed'$$

where the mean vectors of U, V, are zero and their covariance matrices are equal [with diag(R) = diag(I)]:

$$e'U = e'V = 0', \; U'U/N = V'V/N = R$$

For convenience, we divide by N instead of N-1 to define covariances in this exact illustration. With this notation, the mean difference vector is

$$(\bar{x} - \bar{y})' = e'(X - Y)/N = 2d'$$

Since the mean vector $\bar{z}'$ of the 2Nxp pooled score matrix Z is zero and its sample size is 2N, its covariance matrix reduces to

$$C = Z'Z/2N = R + dd'$$

To stress the fact that this matrix can be viewed as the sum of a pxp matrix A of rank one that has been perturbed by adding a "small" matrix of perturbations, $E = R - I$, let us write it

$$C = (dd' + I) + (R - I) = A + E$$

The largest latent root of $A = dd' + I$ is $d'd+1$ and the associated eigenvector is $d$.

As the mean difference vector $2d$ increases, the rank 1 matrix A will increasingly dominate the perturbation matrix E, which remains unchanged. Therefore, if $d$ is chosen large enough, it will approximate the largest eigenvector of the pooled covariance matrix C. They will become virtually collinear as the length of the mean difference vector $2d$ continues to increase, so that their correlation will approach unity, as long the variance of its components does not vanish. A more appropriate measure of the collinearity of $d$ with the largest eigenvector of C (which still works when the variance does vanish) is the cosine, which Jensen calls "the Tucker–Burt coefficient of congruence."

I initially believed that the relation predicted by Spearman's hypothesis depended on the positive manifold, since it implies that the dominant eigenvector of C is unique and positive, as is the black–white mean difference vector. I was mistaken in this. It is now clear that the correlation between the largest eigenvector of $C = R + dd'$ and the mean difference vector $2d$ has

Table 1 (Schönemann). *Numerical simulation demonstrating validity of Spearman's hypothesis regardless of the factor structure of R*

A: p = 5, N = 20. Positive Manifold and positive *d*:

Common correlation marix R randomly generated under the constraints of gramianness and nonnegativity:

| | | | | |
|---|---|---|---|---|
| 1.0 | | | | |
| .78 | 1.0 | | | |
| .56 | .92 | 1.0 | | |
| .84 | .81 | .69 | 1.0 | |
| .50 | .72 | .77 | .82 | 1.0 |

Random vector *f* for defining mean difference vector $2d = 2sf$:

.24 .94 .37 .25 .02

Results of simulation for varying s (max(r)=: largest root of R, max(c)=: largest root of C):

| s | r | cos | max(r) | max(c) | $d'd+1$ |
|---|---|---|---|---|---|
| .2 | .45 | .78 | *3.97* | *3.80* | 1.05 |
| .5 | .52 | .79 | 3.97 | 3.97 | 1.29 |
| 1.0 | .99 | .87 | 3.97 | 4.59 | 2.15 |
| 2.0 | 1.00 | .96 | 3.97 | 7.50 | 5.59 |
| 5.0 | 1.00 | 1.00 | 3.97 | *31.34* | *29.67* |

B: p = 10, N = 50. Signs in R and *d* unconstrained.

Randomly generated R under constraint of gramianness:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1.0 | | | | | | | | | |
| .26 | 1.0 | | | | | | | | |
| −.14 | −.60 | 1.0 | | | | | | | |
| −.13 | .21 | −.51 | 1.0 | | | | | | |
| .20 | .03 | .04 | −.51 | 1.0 | | | | | |
| −.09 | −.10 | −.09 | .11 | .12 | 1.0 | | | | |
| .50 | .15 | .14 | −.48 | .23 | −.52 | 1.0 | | | |
| −.27 | −.24 | .06 | −.25 | −.12 | .27 | .16 | 1.0 | | |
| .38 | −.10 | .05 | −.14 | .07 | −.22 | .43 | −.19 | 1.0 | |
| .52 | .39 | −.33 | −.10 | −.15 | −.27 | .11 | −.01 | .13 | 1.0 |

Random vector *f* used to define $2d = 2sf$:

.68 1.03 −.40 .95 .37 .38 1.17 .69 −.61 −.34

Results of simulation for varying s:

| s | r | cos | max(r) | max(c) | $d'd+1$ |
|---|---|---|---|---|---|
| .2 | .12 | .15 | 2.58 | *2.60* | 1.21 |
| .5 | .60 | .72 | 2.58 | 3.13 | 2.30 |
| 1.0 | .99 | .99 | 2.58 | 6.09 | 6.20 |
| 2.0 | 1.00 | 1.00 | 2.58 | 21.66 | 21.80 |
| 5.0 | 1.00 | 1.00 | 2.58 | *130.91* | *131.04* |

nothing whatsoever to do with the factor structure of the common correlation matrix R because its roots affect only the constant perturbation matrix E = R−I. Hence this correlation cannot shed any light on "the nature of *g*": Spearman's hypothesis will hold regardless of the specific form of R and *d*, provided only the mean difference vector $2d$ is large enough.

To illustrate this concretely, the results of two computer simulations are given in Table 1. In both examples, the matrix R and the vector *f* were randomly drawn but held constant as the scalar s was varied to define successively larger $d = sf$. The score matrices U, V, were not exact in this simulation, and they were randomly redrawn for each run. In the first example, R was positive throughout (Positive Manifold) but did not satisfy Spearman's factor model. In the second example, R contained many negative elements. As can be seen from Table 1, in both cases the largest latent root of C wanders from the largest latent root of R to $d'd+1$, the largest latent root of A, so that the associated eigenvectors become more and more collinear, whether R forms a Positive Manifold or not. The correlations (r) and the cosines (cos) asymptote fairly rapidly. In practice they will probably be somewhat smaller because the assumption of the homogeneity of the covariance matrix will be violated.

3. In his book *Bias in Mental Testing* Jensen has himself offered "an alternative interpretation" of Spearman's hypothesis as an artifact: "whites and blacks differ merely in overall level of performance on all test items . . . and those items (or subtests) that contribute most to the true score variance (by virtue of high reliability and optimal difficulty level) among individuals of either race thereby also show largest mean differences between the races and they are also the most heavily loaded on a general factor (i.e., the first principal component) that, by its mathematical nature, necessarily accounts for more of the variance than any other factor, *regardless of the psychological nature of the first principal component extracted from the particular collection of tests.*" (Jensen 1980a, p. 548f., my emphasis). In other words, since blacks, for whatever reasons, score lower than whites on the "intelligence subtests" that define the first principal component, we in effect select these groups on this principal component, so that tests that correlate with it more strongly will also contribute more strongly to the between-group mean differences. I still regard this as a perfectly adequate explanation in nontechnical terms. However, in his target article, Jensen now rejects his earlier alternative explanation of Spearman's hypotheses as "far too superficial."

Once again I am able to agree with positions Jensen abandoned some time ago, while having to disagree with his more recent, revised positions. This was already the case regarding the probable causes of the undisputed black–white differences on conventional "intelligence tests," which is the theme of the target article. I am still quite comfortable with the explanation Jensen gave in 1966: "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). On the other hand, I find it increasingly more difficult to keep up with his recent forays into factor theory (Schönemann 1981; 1983), which Jensen presumably deems necessary because he now regards his 1966 position as far too superficial.

ACKNOWLEDGMENT:
I would like to thank Professor Michael Drazin, Department of Mathematics, Purdue University, for the crucial hint that this problem may be related to perturbation theory. I would also like to express my gratitude to the Ludwig Maximilian Universität of Munich, Germany, and the Deutsche Forschungsgemeinschaft for partial support of a visiting professorship in Munich.

## The black–white differences are real: Where do we go from here?

Keith E. Stanovich
*Department of Psychology, Oakland University, Rochester, Mich. 48063*

In his survey of the "race–IQ debate," Flynn (1980, pp. 210–11) laid out the alternatives in stark language: (1) the IQ gap is real and is due to environments that cripple cognitive development; (2) the population differences are real, but it is not the environment but genetic factors that handicap blacks; (3) the problem is with the tests themselves. For completeness, we might add a fourth alternative: (4) the differences are real and are a combination of 1 and 2. Both science and social policy are best served if we face the stark alternatives. But, given the minefield that surrounds the alternatives, it pays to proceed cautiously. Thus, I suggest a conservative interpretation of Jensen's target article on Spearman's hypothesis, one that views the paper as a con-

tinuation of the arguments in *Bias in Mental Testing* (henceforth *Bias*). Jensen himself offers this as an alternative: "In other words, the first principal factor of this battery of tests discriminates between black and white individuals on the same basis as it discriminates between individuals in the same population, whether or not the first principal factor is psychologically interpretable as Spearman's *g*. This would be the expected outcome, of course, if the tests in the battery were not biased in discriminating individual differences." In short, the results have implications independent of arguments about the theoretical interpretation of *g*. One implication is that given the evidence in the target article, together with that in *Bias*, the scientifically appropriate course would seem to be to reject alternative 3. [See also *BBS* multiple book review of *Bias* in *BBS* 3(3) 1980.]

In reviewing *Bias*, Scarr (1981a) referred to the "ghosts" that haunted it (e.g., "The apparition of racial genetic inferiority rises from this book," p. 330). Her point was that although the title of the book implied a focus on alternative 3, evaluations of 1 and 2 were frequently implicit. Fortunately, Jensen's present target article is not plagued by ghosts, and this facilitates the evaluation of its scientific contribution. One of the few exceptions occurs when Jensen gives a misleading impression regarding the meaning of the negative correlation between black–white differences and socioeconomic class (SEC) profile in the Jensen and Reynolds (1982) study. His conclusion ("contradicts the notion that the pattern of subtest differences in test performance merely reflects the overall black–white difference in socioeconomic status") may lead to the mistaken inference that SES differences are not related to the zero-order subscale differences. The potential confusion is created because it is not emphasized that the profiles are of correlations with full scale IQ *partialed out*. It is the black–white differences in IQ-partialed abilities – not the overall differences – that are unrelated to SES. In fact, in Jensen's note 4 we learn of .73 and .57 correlations between SES and *g* loadings. Furthermore, in Jensen's previous analysis (*Bias*, pp. 536–37) of a study included in the target article (Nichols 1972) the correlations between the mean black–white difference on a subtest and the SES correlation with that subtest were .72 (whites) and .79 (blacks). In short, tests that were more highly correlated with SES tended to show larger black–white differences, evidence suggesting a linkage between SES and race differences in IQ.

Some will view the discussions of the information-processing correlates of *g* as containing a "ghost" of the genetic argument, but, again, a more conservative interpretation is preferable. We are presented with evidence that *g* is related to physiological variables such as the latency of averaged evoked potentials in the brain and to elementary cognitive tasks that are "related to physiological processes at the interface between brain and behavior": in short, that *g* (a composite of performance on many cognitive tasks) is related to brain processes. If one believes that our current intelligence tests engage at least some nonspecific mental operations and that mental processes reflect brain processes (conceded by all but the wildest of dualists) then such a correlation should be no surprise. The existence of the correlation does not require the acceptance of any theory of *g* beyond its operational definition in terms of the factor analysis of sets of mental tasks. It does, however, reinforce the argument that differences both within and between populations are *real* and not the result of test bias or peculiar item selection; that is, it reinforces the decision to reject hypothesis 3. It is important to note that the finding is moot regarding the causes of the differences.

Where do we go from here, then? No doubt Jensen will continue to pursue a research program focused on a theoretical understanding of *g*. Other researchers, while not denying the empirical facts that are the basis of the *g* construct, are not so enamored of it as a basis of theory construction, and they will pursue other alternatives (e.g., Detterman 1982; Sternberg 1984b). My guess is that the reaction-time (RT) research program of Jensen's will have some success but will meet some vexing problems when it eventually addresses the questions of the malleability, stability, and trainability of the elementary task performances that are the center of the method. For example, there are already reports of some rather large training and feedback effects on the RTs of moderately and severely retarded adults (Wade, Hoover & Newell 1984).

Accepting the falsification of hypothesis 3 will allow researchers to focus their attention on the other hypotheses, something many have of course been doing for years. Others may wish to ignore the group-differences issue and attack the problem of training academic and cognitive task performance. Either way, accepting the implications of Jensen's evidence will speed the cumulative growth of knowledge in a way that rarely occurs when falsified hypotheses are allowed to live on to clutter the intellectual landscape.

Researchers who pursue the group-differences issue will find a problem much less tractable than the evaluation of hypothesis 3. Nevertheless, some progress has been made. Some relatively recent developments include progress on quantifying important home environment variables (Bradley & Caldwell 1984; Durkin 1982; Price 1984; Sameroff & Seifer 1983; Thomas 1984), discovery of cultural and social differences between black and white families of the same SES (Blau 1981; Trotman 1977; Tulkin 1968), the development of some preliminary theories of cultural differences that might explain black–white differences in test performance (Flynn 1980; Laosa 1982; Ogbu 1982), studies that attempt to separate genetic background from the effects of home environment (Plomin 1983; Scarr 1981b; Schiff, Duyme, Dumaret & Tomkiewicz 1982; Wilson & Matheny 1983), studies of cumulative cognitive deficit (Jensen 1977b), and theoretical accounts of the cognitive differences between different populations (Borkowski & Krause 1983; Feuerstein 1979).

Perhaps more relevant to the academic achievement problems of all children will be work in cognitive training. Researchers have begun to turn from the tired debates about "training intelligence" to focus more on training performance on tasks with direct academic relevance such as arithmetic skills and reading. Developments in the latter field may be of general interest to intelligence researchers because in the past ten years many different techniques for intervening to facilitate reading have been developed. Researchers in this area have been more concerned with raising the performance levels of all children than with reducing or explaining population differences, and this focus seems to have paid off, because many different types of children seem to benefit from the training techniques. For example, Williams (1980) field tested a low-cost structured program of phoneme analysis and blending on low-IQ children in several Title I classrooms and found significant long-term increases in decoding skills. Hansen and Pearson (1983) found that training in inferencing benefited the reading comprehension of below-average fourth-graders more than above-average fourth-graders. Bradley and Bryant (1983) found that training preschoolers in sound categorization led to gains in reaching achievement three years later. Many other examples of successful training programs exist, some focusing on decoding (e.g., Blanchard 1980; Samuels 1979; Wallach & Wallach 1979) and some on comprehension (see Ryan 1981; Tierney & Cunningham 1984).

In short, while the furious debate about the fairness of IQ tests has raged, reading researchers have been slowly and painstakingly developing ways of remediating deficits in the very skills that the tests were designed to predict. In the process, some theoretically interesting things about cognitive development have also been learned (Pearson 1984). The demise of the debate about testing can only further the practical and theoretical progress that has already been made. Jensen has contributed to the cumulative progress in the testing and individual differences fields with *Bias* and this target article. Now, let's move on.

## The black–white differences and Spearman's *g*: Old wine in new bottles that still doesn't taste good

Robert J. Sternberg

*Department of Psychology, Yale University, New Haven, Conn. 06520*

So what? As Jensen himself points out, we have known since the early 1900s that blacks score lower than whites on conventional intelligence tests. If there is an interesting research question lurking in this finding, it is why there is such a score difference. His target article does not advance our knowledge, but merely restates what we already know in repackaged form. Consider the following facts.

First, unrotated principal-component or principal-axis factorial solutions tend to give a strong general factor followed by a succession of weaker bipolar factors. This is a statistical certainty that has nothing to do with psychology. Hence, if one factor analyzes a series of mental-ability tests and extracts an unrotated solution, one will obtain a relatively strong factor followed by weaker and less reliable group or specific factors.

Second, differences in scores on tests that are so factor-analyzed will be due primarily to individual differences in scores on the general factor. After all, it is primarily the first factor (in the unrotated solution) that the tests measure. With rotation, the variance in the scores will be distributed more, so that the loci of differences may be broadened. But it would be extremely odd, in an unrotated factorial solution, if most or even much of the variation in scores obtained were due to factors other than the strongest factor, namely, the general one.

Third, Jensen's analysis merely confirms the statistical near-certainties addressed by the first two points above. The high correlation of score differences with *g* is almost a restatement of the fact that blacks score lower than whites on conventional intelligence tests. The same relations would hold for virtually any other attribute that might be measured and then subjected to a principal-components or principal-axis solution. One would scarcely expect the main locus of differences to be in the weaker and less reliable factors. Consider, for example, multiple indirect measures of body weight, such as amount of body fat, amount of fluids in the body, girth at the waist, and so on. (I use indirect measures, because intelligence, unlike weight, cannot be directly measured.) Suppose one were to factor-analyze such measures and extract a first principal component or factor. It would scarcely be surprising if some index correlated with measured differences in obesity showed its greatest correlation with the first principal component or factor obtained from measures such as those named above. But such a correlation would tell us nothing about (a) the various antecedents of obesity, (b) why some people tend to be more obese than others, (c) what can be done to remedy obesity, or, most importantly, (d) why the correlation is interesting in the first place. Note that not even a true measure of differences in obesity – scale weight – would address any of these questions.

My point is simply this. Jensen's analysis merely restates in a more complicated way what has already been known for a long time: Blacks score lower than whites on conventional intelligence tests. As Jensen also notes, they score lower on some other measures as well, which are correlated with conventional IQ. But Jensen's analysis answers none of the more interesting and timely questions, such as why the score difference holds, what can be done to remedy it, or why the difference matters in the first place. At best, Jensen's attempts to interpret the data – which I do not regard as major can only give comfort to those who would like nothing better than to hear the explicit message that an important practical implication of the results is that blacks will have a greater handicap in the educational, occupational, and military assignments that are most highly correlated with measures of general intelligence. If that is the best one can do by way of conclusions, then the minimal scientific gain of these data is more than offset by the potential loss to society from such interpretations of the data.

This last issue brings me to my final point. [It is an *ad hominem* point and has only been included with the permission of the author, ed.] Jensen is a competent scientist and scholar. But another major criterion by which scientists are judged is their choice in scientific problems. Scientists are judged at least as much by the nature of the problems they elect to study as by the ways in which they go about solving them. Jensen's investigations into the nature of intelligence show that he can select problems well and address them well. Although I do not agree with his theory of the nature of intelligence, I have no difficulty in respecting the theory and the research behind it. But I am at a loss as to why Jensen persists in studying the problem of black–white differences. Despite my own distaste for the problem, I might be impressed by research that helped us understand the causes of these differences or what could be done about them. But Jensen's research has not illuminated any of these more difficult and scientifically interesting issues; rather, it has merely restated the same finding again and again, albeit in a slightly different form each time.

I suppose that other scientists too have their preoccupations. I only wish Jensen would make better use of his considerable talents. I hope he is remembered for his basic and scientifically interesting research on the nature of intelligence and not for his derivative research on black–white differences. I fear that this will not be the case.

## Interpretation of black–white differences in *g*

Philip E. Vernon

*Department of Educational Psychology, University of Calgary, Calgary, Alberta, Canada T2N 1N4*

As we have come to expect of Arthur Jensen, the clarity of his arguments and the manner in which he seeks to plug all possible loopholes make a very persuasive case for what is likely to be an unpopular view of black–white cognitive differences. There are some ambiguities to discuss, however.

The introductory pages emphasize the invariance of *g*, regardless of what tests are factorized. But since Jensen uses what is essentially the total of (standardized) scores on all tests as his criterion of *g*, the content of his measure can vary considerably with different choices of tests. Thus, Wechsler's correlation of total Verbal with total Performance tests on WISC-R is only .67 (which hardly bears out Jensen's claim of a *r* of .80 between the *g*'s given by the two batteries). Several writers have criticized Jensen's use of the first unrotated principal component or factor as his measure of *g*, since this is certainly not invariant from one battery to another. However, further on in the article, Jensen admits this weakness, and claims, justifiably, that his *g* is quite highly invariant provided that (1) the investigator uses a large number of varied tests, (2) some of the tests aim to measure Spearman's eduction of relations and abstract thinking and (3) these obtain the highest loadings on the first factor. This seems to me a sufficient definition of *g* for working purposes, but it does imply that the measure of *g* depends to some extent on the subjective judgments of the factorist.

Second, we should accept that ECTs (elementary cognitive tasks) such as choice reaction time, inspection times, and EEG evoked potentials show stronger positive correlations with a *g* (as defined above) than previously believed. Jensen's explanations of this finding vary, however. He refers to ECTs as measuring speed and efficiency of cognitive processing, but elsewhere he refers to them as the capacity of the working memory. The latter explanation could scarcely account for *g*'s higher correlations with choice reaction time than simple reaction time. Eysenck (1982b) emphasizes yet another aspect,

namely, freedom from error in the transmission of neural signals. But elsewhere in his paper Jensen still writes of *g* as reflecting "the complexity of the mental processes required for a task," which is quite in line with contemporary accounts of the nature of intelligence. But how can this be reconciled with Jensen's and Eysenck's claims for the large *g* variance of very simple cognitive tasks, and even for neuropsychological measures? For myself, I could accept that a combination of a variety of ECTs would correlate up to .50 with *g* in a representative sample of adults or older children, that is, 25% of the variance; but not that correlations would be in the .70s and over. For these would imply that ECTs are better measures of *g* than other cognitive tests that are much more complex but whose *g* loadings are only moderate (e.g., word reading, clerical speed and accuracy, or number tests). We badly need an investigation with a representative population, one that would show the relative loadings of evoked potentials, ECTs, moderately complex cognitive tests and would use highly complex verbal and nonverbal reasoning tests as our criterion of *g*.

There is another possible explanation that has been ignored by advocates of ECTs. The existence of a correlation between such tests and *g* does not demonstrate that the efficiency of brain processing is, to some extent, the cause of human intelligence. Why should not environmental stimulation, which, according to Hebb (1949), is essential for mental growth, also improve the underlying brain mechanisms? Thus Krech, Rosenzweig and Bennett (1962) have demonstrated that stimulation of baby rats by frequent handling not only improved their later maze learning but also brought about anatomical and biochemical changes in the brain. If this applies to man, the correlation between simple neuropsychological or elementary cognitive tests and *g* could be partly attributed to mental stimulation and growth affecting the brain.

Third, Jensen is careful (unlike some others) not to claim that ECTs represent genetic differences that would imply that the black–white difference was, at least in part, a biological race difference. Yet some of his opponents are unfortunately, only too likely to regard the target article as another manifestation of racism. Jensen has never denied that environmental factors play an important part in determining the phenotype of *g*, though he has estimated the environmental variance to be as low as 20%, whereas psychologists with middle-of-the-road views would probably say 40–50%. He has indeed attacked "X-factors," that is, hypothetical environmental influences that are claimed, without scientific evidence, to handicap black children. But there are a number of adverse conditions that have been validated, for example, malnutrition, poor maternal health, and the type of verbal interactions between mother and child. True, such factors tend to be positively correlated, so that their combined influence might be less than expected. We are still far too ignorant of the crucial dimensions of environmental stimulation. However, the present article would be less liable to misinterpretation had the author drawn attention, yet again, to the interactive conception of *g*.

## Focusing on trainable *g*

Arthur Whimbey

*3051 S. Atlantic Avenue #503, Daytona Beach Shores, Fla. 32018*

Professor Jensen's paper is one in a series of articles he has recently written emphasizing that there is a measurable human capacity, called *g* for general intelligence, that not only influences test scores but also significantly affects professional performance (Jensen 1984a) and therefore warrants major research efforts. Jensen's target article argues that there is a difference of one standard deviation between blacks and whites in *g*, and that basic neural processes, reflected in simple and complex reaction time, play an important role. To balance the picture, this commentary will focus on some ideas briefly noted in the last paragraphs of Jensen's paper.

The fourth paragraph from the end contains this statement: "As the present black–white difference in general speed of processing is only about one-third as large as the mean black–white difference on the ASVAB, it seems likely that the *g* of the ASVAB (and similar achievement-oriented psychometric tests) also involves types of higher-order processing other than the quite elementary processes measured by the present tasks." The next-to-last paragraph continues: "It seems likely that the 'software' components of intelligent behavior (the so-called metaprocesses of executive control, problem-solving strategies, predicting and monitoring one's own performance, and the like) may be more readily trainable than the "hardware" components (speed of encoding, short-term memory capacity, retrieval of information in long-term memory, etc.) . . . . We are even uncertain to what extent these hardware components of human information processing are amenable to special training."

A number of researchers have addressed the problem of improving *g* through training (reviewed in Whimbey 1975). By asking high- and low-*g* students to think aloud while solving problems, a consistent difference in processing style was discovered. This can be illustrated with the following verbal analogy, a type of problem included on several tests that are highly loaded in *g*, such as the SAT–Verbal listed in Jensen's Table 2.

Elephant is to small as ———is to ———.

| | |
|---|---|
| (A) large: little | (C) lion: timid |
| (B) hippopotamus: mouse | (D) turtle: slow |

Low-*g* students miss this analogy because they do not explicate the relationships between the pairs of words with enough detail and precision. They frequently pick alternative (A) and, when asked to explain their choice, answer that "an elephant is not small and large is not little." Or they may say that "elephant and small are opposites, and large and little are also opposites." They have been characterized as "one-shot thinkers" by researchers because they tend to jump to an answer without a sufficient step-by-step analysis. They do not spell out the relationship that an elephant is an animal and that smallness is a quality which is not characteristic of that animal. And they do not spell out the relationships between all the other pairs of words until the correct answer (C) is found. Furthermore, their one-shot thinking is a habitual way of responding, extending to mathematical and figural as well as verbal problems. Jensen notes that Spearman defined *g* as ability in "educing relationships." Our research indicates that high-*g* students have learned to engage in more mental processing in order to educe correct relationships.

To improve the analytical reasoning of students, a procedure has been devised called Thinking Aloud Pair Problem Solving (TAPS), in which pairs of students take turns thinking aloud as they solve problems (Whimbey & Lochhead 1982). The acronym TAPS reflects that the procedure *taps* mental processing, bringing it out into the open where it is available for observation, guidance, and feedback.

One program that has been using TAPS for several years is Project SOAR at Xavier University, a traditionally black college in New Orleans. SOAR is a prefreshman program providing students with about 40 hours of training in analytical thinking. For students whose initial SAT score is below 700, the average gain is about 110 points (Hunter et al. 1982). This is a little more than one-half a standard deviation, which may be compared to the one-standard-deviation difference between blacks and whites noted by Jensen. Furthermore, SOAR students are twice as likely as other Xavier students to pass their freshman science and math courses, which suggests that thinking ability (*g*), not just test scores, has been improved.

Recently TAPS has been incorporated into the teaching of reading comprehension (Whimbey 1984) because reading com-

prehension ability is highly loaded with *g* and is widely (but not well) taught in the schools. As an illustration, low-*g* students are initially unable to correctly answer questions such as the following.

> In geology the last 11,000 years are called the Recent epoch, and the Recent epoch together with the Pleistocene epoch makes up the Quaternary period. Moreover, the Quaternary together with the Tertiary period makes up the Cenozoic era. The Cenozoic is the only era in which periods are broken down in epochs. The other eras are subdivided only into periods. The era immediately preceding the Cenozoic is the Mesozoic, during which the Jurassic period represents the age of the dinosaurs, although these giant reptiles appeared before the Jurassic and became extinct later than the Jurassic – in the Triassic and Cretaceous periods, respectively. In the still earlier Paleozoic era the first sharks and reptiles appeared during the next-to-last period, the Carboniferous, while in the last period of this era, the Permian, reptiles flourished. Preceding the Carboniferous period was the Devonian, and before that, from earliest to latest, the Cambrian, Ordovician, and Silurian periods. Write the 11 periods in order from earliest to latest on a diagram. Do not write eras or epochs.

However, their performance improves greatly after using TAPS while working through a series of 60 problems, beginning with easy ones like this (Whimbey 1983).

> Atlanta has a larger population than Birmingham but a smaller population than Chicago. Write the names of the three cities in order on the diagram.

larger population

smaller population

Significant gains have been made on a standard reading measure, the Iowa Silent Reading Test, but an evaluation of the practical, long-term impact will take several more years (Whimbey 1981).

In closing I would like to draw attention to a few additional research questions raised by Jensen's findings. If blacks have slower reaction times, how have they come to dominate boxing and excel in other sports like baseball? If they can't get their finger off the button of the reaction-time apparatus as quickly as their white counterparts, how do they duck their punches and hit their pitches so well? The answer is not simply superior muscular strength, because both American and international weightlifting is dominated by whites. Nor is it muscular coordination, since Jensen's Figure 6 shows no difference between blacks and whites here. Aside from athletics, the reaction-time research seems at odds with the prominence of blacks in the creation and performance of jazz, some of which (for example, that of Thelonius Monk) is rich, complex, and sophisticated. As Jensen suggests, much research is still needed on *g* and other human abilities.

## Jensen's support for Spearman's hypothesis is support for a circular argument

James R. Wilson

*Institute for Behavioral Genetics, University of Colorado, Boulder, Colo. 80309*

Having received an earlier draft of the target article, we prepared and submitted a paper commenting on several aspects of it, and we included new analyses from the Hawaii Family Study of Cognition that were relevant to the argument (Nagoshi et al., in press). We furnished Professor Jensen with a copy of this manuscript; however we unfortunately see no indication in his article that he has considered the arguments or data presented therein.

I would like to reiterate here one of the arguments we presented: "Because a group difference on *g* requires group differences on tests which load on *g*, an observed group difference in general mental ability may necessarily result in a correlation between group differences on individual tests and their *g* loadings" (Nagoshi et al., in press). Another way of saying this is that, given a substantial group difference on *g* (such as is commonly reported for blacks vs. whites), it is hardly surprising that there will be a substantial group difference on those tests which load most heavily on *g*, since they in a very real sense *define g*. Whether it is *g* that we conceive to be theoretically prior or the actual tests hardly matters; we have but one phenomenon (the group difference), and we add nothing to our understanding of the phenomenon by running the argument around in a circle.

# Author's Response

## The black–white difference in *g*: A phenomenon in search of a theory

Arthur R. Jensen

*School of Education, University of California, Berkeley, Calif. 94720*

The 29 commentaries present such a diversity of opinions and observations on so many different aspects of the target article as to make it virtually impossible to do justice in my response to every single point. It will be necessary to focus on those issues that show some communality among commentators or that raise questions that seem most central to the main findings and are most apt to help clear the way for further research and theoretical formulations. I will try, however, to touch upon as many of the points raised by the commentators as possible, even if it means adopting a fairly telegraphic style, with abrupt changes of topic.

Most of the comments fall into one of two main categories: (1) Spearman's hypothesis per se and the psychometric and statistical problems surrounding it, and (2) the relation of response latency, or reaction time (RT), on a variety of elementary cognitive tasks (ECTs) to psychometric *g* and associated methodological and interpretive issues.

### Spearman's hypothesis per se

There is rather less explicit agreement or disagreement with Spearman's hypothesis than I had expected, given that it was the central theme of my target article. The test of Spearman's hypothesis is the significant and consistent (across 12 studies) correlation between psychometric tests' *g* loadings and the magnitudes of the mean black–white differences on the tests (expressed in standard score or $\sigma$ units). Nine of the commentators (**Brand, Cattell, Eysenck, Gordon, Jones, Kline, Nettelbeck, Nichols,** and **Stanovich**) explicitly regard the hypothesis as having been borne out by the evidence. Two (**Gustafsson** and **Baron**) express doubts or propose a counterhypothesis. Three (**Johnson & Nagoshi, Schönemann,** and **Wilson**) seem to accept the hypothesis as borne out,

but claim that this outcome was inescapable, being a mathematical artifact or a "circular argument" preordained by the workings of factor analysis. The remaining 13 commentators express no opinion one way or the other regarding Spearman's hypothesis per se.

**Sternberg,** however, does not quite fall into any of these categories. He claims that the analysis merely restates what we already knew, namely, that blacks score lower than whites on conventional intelligence tests. That fact has indeed been well known for a long time. But that is not the issue specifically addressed by Spearman's hypothesis, which arose in the first place from the observation that there is considerably more to the black–white difference on psychometric tests than just the overall difference itself, the important point being that the magnitude of the black–white difference *varies* across different tests. I have not found any thorough examination of this phenomenon anywhere in the previous literature. There would be little if any scientific leverage in observing still one more black–white difference on one more test. What has not been discussed, much less understood, is the *variation* of differences, which, if it proves to be a reliable phenomenon, could provide some leverage for further understanding the nature of the black–white differences in cognitive performance. Testing Spearman's hypothesis, or investigating the variation among differences, is an essential step toward an adequate account of the black–white differences on psychometric tests. Sternberg's belittling of this aim is surprisingly unanalytical for an otherwise generally very analytical psychologist. Attempting to reduce these findings to nothing more than the well known average difference of about 1 $\sigma$ on "conventional intelligence tests" not only misses the essential question that gave rise to Spearman's hypothesis, but also tars such research with the popular opprobrium attached to IQ tests. Is it not a reasonable question to ask (assuming we are interested in the subject at all) which content features or psychometric characteristics of tests are associated with the conspicuous variation in the size of the mean black–white difference on different tests? Might not such inquiry afford clues as to the essential nature of the black–white difference, or at least point investigators in the best direction for further study? What the present analysis consistently shows is that variation in the black–white difference is not systematically associated with such surface or content characteristics of tests as whether they are verbal or nonverbal, culture-loaded or culture-reduced, performance or paper-and-pencil, pictorial or figural, and so on but is most consistently associated with a latent trait, *g*, or the largest common factor in virtually any sizable battery of diverse cognitive tasks. The nature of the black–white difference, therefore, must be sought in the nature of *g* rather than in the intellectual content and other surface features of conventional psychometric tests.

**The inevitability–circularity–artificiality claim.** Several commentators regard the outcome of testing Spearman's hypothesis as inevitable or artifactual or a circular argument. **Jones** believes that any other conclusion from the results would be totally unexpected. If it is unexpected to Jones, it is largely because Jones, a sophisticated psychometrician who has investigated black–white differences, already knows the kinds of tests that show the largest differences and the fact that constructors of conventional intelligence tests select item types that are *g*-loaded. They select such items not necessarily because they are *g*-loaded but because it is found that the most *g*-loaded items maximize predictive validity for the kinds of practical criteria for which tests are commonly used. The literature on group differences in test scores attributes differences almost exclusively to specific contents and surface features of tests, and the demonstration of what Jones regards as totally expected (i.e., the substantiation of Spearman's hypothesis) actually contradicts the conventional and popular view of black–white test differences. Moreover, not all group differences on a battery of psychometric tests are *g* differences, as I showed in the comparison of preverbally deaf children and normal-hearing children. The correlation between WISC-R subtest group differences and subtest *g* loadings for deaf and hearing children was in fact negative–the opposite of the black–white comparison. Jones assumes that this "unexpected" finding must be due to a different pattern of *g* loadings for deaf and hearing children. Yet Braden (1984, p. 406) has reported a congruence coefficient of +0.988 between the *g* factor loadings of the deaf and hearing groups, that is, virtual identity of the *g* factor across these groups. But the profile of group differences on the subtests is negatively correlated with the profile of the subtests' *g* loadings. True, the overall hearing–deaf difference is only about one-fifth as large as the typical black–white difference. But that cannot be the cause of this outcome. I have shown that the effect of inbreeding depression is to lower the WISC IQ just about as much as Braden reported for the effect of deafness on the Performance IQ. Yet the varying effects of inbreeding depression on the WISC subtests are correlated about +0.80 with the subtests' *g* loadings (Jensen 1983a).

As for **Wilson**'s claim of circularity, it is his own agrument, not Spearman's hypothesis, that is circular. Of course, if one postulates (as does Wilson) that a group difference is mainly a *g* difference, then it is indeed inevitable that the group differences on various tests will be correlated with the tests' *g* loadings. One can always make a proposition circular by stating the conclusion in the premises. The same fallacy is voiced by **Johnson & Nagoshi,** who, in their first sentence, state that "any group difference in *g* would of necessity be reflected in the tests that load on *g*." This is of course a mere tautology. Change the statement to "any group difference in IQ (or total score, etc.)" and it is no longer a tautology or inevitability. After stating the tautology, Johnson & Nagoshi claim that "his finding in itself casts serious doubts on the validity of Jensen's conclusions concerning black–white differences in cognitive abilities." But this claim is a non sequitur. Do Johnson & Nagoshi mean to imply that this tautology contradicts Spearman's hypothesis? After their puzzling first paragraph, Johnson & Nagoshi go on to show some other theoretically interesting relations between *g* and certain familial and social variables in their own study of various populations in Hawaii, and one could hardly disagree with their concluding statement that "there is clearly a need for even more basic research on the nature of *g*."

**Schönemann** illustrates the same kind of tautology mathematically, showing that if one "builds in" a large enough difference between groups on a number of corre-

lated variables (which thereby yield a general factor), the groups will differ on the general factor. It appears to me that this is another case of stating the premises or conditions necessary for a given outcome. I do not see that it differs essentially from saying, for example, that if two cars start a race from the same point and traverse the same distance, the car with the faster average speed will cross the finish line ahead of the car with the slower average speed. But do the premises that the average speeds are different and that the distance is the same make the observation that one car arrives at the finish line ahead of the other merely an artifact or an illusion? On the other hand, one can point to many conjectures by psychologists in the literature on black–white IQ differences that are contradicted by the very conditions or premises that Schönemann demonstrates as sufficient for Spearman's hypothesis, for example, equal covariance matrices and large enough differences on the mean vectors. But Schönemann's demonstration apparently leads him to agree with a false, or at best theoretically too limited, conclusion, namely, the statement he quotes from my *Bias in Mental Testing* (1980a). Although the conditions stated therein *could* produce the appearance of Spearman's hypothesis, these conditions are neither necessary nor sufficient to account for the actual findings. Since 1980, I have explicitly investigated this matter, and I find that neither the variation in the *g* factor nor the varying magnitude of the black–white difference on various tests is at all dependent on differences in test reliability or on variation in item or subtest difficulty level. High and low *g*-loaded tests, even when perfectly matched on reliability, still show large and small black–white differences, respectively. Moreover, we have found that different single items of the Raven Progressive Matrices test can differ in their *g* loadings even when they are perfectly matched on item variance [i.e., $p(1 - p)$, where $p$ is proportion passing]; the more complex (hence more difficult) items are generally the more *g*-loaded, even when $p(1 - p)$ is the same for the simple and complex items (e.g., $p = .80$ and $p = .20$). In brief, *g* can vary independently of reliability and range restriction, even among tests or items that are quite homogeneous in form and content. It has become increasingly clear in recent years that neither *g* nor the black–white difference on cognitive tests is merely a psychometric artifact.

**Baron** is right in noting that the reliability of a test can affect both its *g* loading and its power to discriminate groups. But this does not mean that Spearman's hypothesis depends on differences in test reliability, although such differences could conceivably simulate an outcome consistent with the hypothesis when the hypothesis was actually false. However, the present results cannot be explained in this way, as I have already shown in the target article. When the *g* loadings and black–white differences ($\bar{D}$) are corrected for attenuation, Spearman's hypothesis still holds (see Table 3 in the target article). The lowering (by about .10) of the correlations between *g* and $\bar{D}$ is adequately explained by the greater restriction of range of the disattenuated *g* loadings. The use of parallel-forms test–retest reliabilities rather than internal-consistency (split-half or K-R [Kuder-Richardson] 20) reliabilities would be a nice addition but would be most unlikely to alter the results appreciably. Although the two forms of reliability are clearly distinct conceptually, empirically the resulting reliability coefficients ($r_{xx}$) generally run quite parallel. I think that Baron makes too much of the partial correlation, in which $r_{xx}$ is partialed out of the zero-order correlation between *g* and $\bar{D}$. Although the resulting partial correlation is capable of testing the null hypothesis, beyond that, its actual magnitude, unlike the correction for attenuation, cannot be interpreted as yielding a closer approximation to the true correlation between *g* and $\bar{D}$. There is, of course, no demonstration of an inherent theoretical connection between a test's parallel-form retest reliability and either its true (i.e., disattenuated) *g* loading or its true discriminability between populations. Reliability and *g* are certainly not the same construct, even though in some test batteries they may be adventitiously correlated. Reliability is largely a function of test length. The Digit Span subtest of the Wechsler scales, for example, has one of the lowest reliabilities in the Wechsler battery and also one of the lowest *g* loadings, and it shows one of the smallest black–white differences of any subtest. In a number of my previous studies, however, I have used repeated parallel forms of the forward Digit Span test to increase the reliability of the composite Digit Span test up to values above .90, that is, as high as the reliability of the Full Scale IQ. Even so, the *g* loading of Digit Span is still much lower than the *g* loadings of, say, Vocabulary and Block Design. Also, the size of the black–white difference on the highly reliable forward Digit Span test is still among the smallest of the differences on any of the many tests we have used in our research (e.g., Jensen 1971; 1973b; 1974a; Jensen & Figueroa 1975; Jensen & Innouye 1980).

**Gordon** has made a striking contribution to the methodology of testing Spearman's hypothesis, based on the equivalence of the congruence coefficient (or index of factor similarity) and the correlation between factor scores. The point biserial correlation ($r_{pb}$) of test scores with the black–white dichotomy is clearly equivalent to the tests' loadings on a black–white factor. The question then is whether factor scores based on this black–white factor were computed for every subject, and if factor scores based on the *g* factor (the first principal component) of all the tests in a given battery were computed for every subject, the correlation between the two sets of factor scores—the black–white factor scores and the *g* factor scores—would be equal to the coefficient of congruence between the tests' loadings on the black–white factor and the loadings on the *g* factor. (Although this equivalence would hold exactly only for a *g* factor computed as the first principal component, and the present analyses are based on the first principal factor or on the Schmid-Leiman second-order *g*, these are only negligibly different from the first principal component in the present data sets. Therefore, Gordon's figures would probably differ only in the third decimal place.) The congruence coefficients shown in Gordon's Table 1 range between .915 and .993, with an average of about 0.97, that is, an almost perfect correlation between factor scores based on *g* and the magnitude of the black–white difference, as Gordon concludes. This is a striking substantiation of Spearman's hypothesis, albeit an inferential substantiation, based on the correctness of Gorsuch's (1974, p. 253) claim of equivalence between the principal component factor score correlation and the congruence coefficient. For those who would like to see a precise

empirical demonstration of this outcome as well, the total Wechsler (WISC-R) standardization data from the study by Jensen and Reynolds (1982) are available and can be subjected to a direct determination of the equivalence of the factor score correlation and the congruence coefficient. This analysis will be done as soon as feasible, and a note on the results will be submitted to a forthcoming Continuing Commentary section in this journal.

**Factor analysis and the nature of *g*.** In the target article I tried to treat *g* as empirically as possible, without bringing in any particular theory of *g* or allowing subjective judgments or theoretical preconceptions to determine the *g* factor or its relation to the black–white difference. As a starting point, I thought it best to take whatever *g* the available data sets yielded by an objective method of analysis, even though some of the available test batteries were rather far from representing an ideal sampling of the whole domain of abilities measured by psychometric tests. Never was a test battery included or excluded because of how well the particular collection of tests conformed to any particular theoretical conception of the "ideal *g*," whatever that might mean. I agree with the observation of **Gustafsson, Jones,** and **Vernon** that the *g* factors extracted from these 11 quite diverse test batteries are bound to vary to some degree, which cannot be precisely determined from these data. As correctly noted by **Kline,** however, the fact that the *g* factor varies somewhat according to the different compositions of these batteries could only attenuate the test of Spearman's hypothesis. Yet the hypothesis was borne out by every battery. **Gustafsson** notes that generally in these particular batteries the tests with the largest *g* loadings and largest black–white differences are of the achievement-laden type frequently characterized as crystalized *g*, or $g_c$, as contrasted with fluid *g*, or $g_f$. But it might well be that in culturally or educationally homogeneous populations (as indicated, for example, by their high similarity in factor structure), verbal and achievement-type tests yield even better measures of $g_f$ than the often less reliable and spatially loaded tests most commonly used to represent $g_f$. The *g* of most of the test batteries used in this study is undoubtedly some amalgam of $g_c$ and $g_f$. But if these batteries could be subjected to a hierarchical or Schmid-Leiman factor analysis along with a much larger collection of tests that sampled more widely the entire psychometric domain, I think it would be a safe prediction that the topmost *g* of the hierarchy (call it Spearman's *g*) would be larger (in variance accounted for) than $g_c$ or $g_f$ or the two combined and that the residualized $g_f$ would be reduced to practically nil, most of it being absorbed by Spearman's *g*. Recent hierarchical factor analyses of test batteries with broad samplings of abilities have shown exactly this picture (Gustafsson 1984; Undheim 1981a; 1981b; 1981c). Spearman's *g* and $g_f$ are either very similar or the same, and much of the variance of the kinds of tests that are usually most heavily loaded on $g_c$ is absorbed into the top hierarchical *g* when residualized by the Schmid-Leiman procedure. Hence one cannot accept as a cogent criticism Gustaffson's comment that my analysis leaves Spearman's hypothesis largely uninvestigated. However, it would be very desirable to see Spearman's hypothesis tested using the broad sample of tests that, in Gustafsson's (1984) own study, yielded what he might consider an "ideal" *g* and led him to conclude that *g* is identical with $g_f$. This conclusion led Gustafsson (1984) to a most important observation: "Formulated in simple terms this result implies that scores obtained on a test consisting of the broadest and most representative sample of tasks are virtually perfectly correlated with scores obtained on a small set of $g_f$ tasks. The most interesting question must then be why the $g_f$ tests have such power of indexing general intelligence" (p. 195).

This, I think, is the most telling criticism of Humphreys's purely descriptive definition of general intelligence, a conception that **Jones** seems to advocate that I should adopt. Humphreys (1971) has defined general intelligence as follows:

> Intelligence is defined as the entire repertoire of acquired skills, knowledge, learning sets, and generalization tendencies considered intellectual in nature that are available at any one period of time. An intelligence test contains items that sample the totality of such acquisitions. The definition of intelligence here proposed would be circular as a function of the use of intellectual if it were not for the fact that there is a consensus among psychologists as to the kinds of behaviors that are labeled intellectual. Thus, the Stanford-Binet and the Wechsler tests can be considered examples of this consensus and define the consensus. (Pp. 31–32)

My own reservations about this definition have been expressed in detail elsewhere (Jensen 1984c). The definition is essentially theoretically barren. In relation to the earlier quotation by Gustafsson, it is a theoretically crucial fact that intelligence, as defined by Humphreys, can actually be measured adequately by a limited number of tests that involve much less than the totality of the repertoire of acquired skills described by Humphreys. One does not need to sample from the totality of this repertoire in order to measure its general factor. In fact, it is now beginning to appear that one may need to measure only certain aspects of the averaged electrical potentials of the brain elicited by auditory "clicks" (Hendrickson & Hendrickson 1980). Humphreys's definition deals only with what **Eysenck,** following Hebb, has termed Intelligence B, which comprises the multifarious manifestations of Intelligence A, characterized by Eysenck as a "capacity of the central nervous system and cortex to process information correctly and without error." There is nothing in the Humphreys definition that would lead one to expect the existence of a *g* factor in the varied repertoire described by his definition or to imagine that the same *g* factor could be measured by tests tapping very different contents of the repertoire–the important phenomenon referred to by Spearman (1927) as "the indifference of the indicator" of *g*.

As an example of this phenomenon, I cited the fact that the *g* factors extracted separately from the Wechsler verbal subtests and the performance subtests are correlated .80 with each other, despite the highly dissimilar contents of the verbal and performance tests. **Vernon** appears to cast doubt on this claim by citing a correlation of .67 between the Verbal and Performance IQs of the Wechsler Intelligence Scale for Children–Revised (WISC-R). I haven't determined the correlation between the *g* factors of the Verbal and Performance subtests of the WISC-R; my statement was based on this

determination for the Wechsler Adult Intelligence Scale (WAIS), in which even the simple correlations between the Verbal and Performance IQs range between .77 and .81 for various age groups (Matarazzo 1972, p. 243). Clearly, very dissimilar test batteries yield very similar *g*s. Is the *g* of all the Wechsler subscales mainly $g_c$, as **Cattell**'s statement suggests, or does it also represent $g_f$ to a substantial degree? One might predict from Gustafsson's (1984) observations that a *g* extracted from such a diverse battery as the Wechsler would most probably come close to Cattell's $g_f$. Raven's Matrices, like Cattell's Culture Fair Tests of $g_f$, is generally considered a quintessential test of $g_f$. It is therefore noteworthy that when the Raven Matrices (Advanced) was factor-analyzed among the 12 WAIS subtests, it showed a higher *g* loading (+0.80) than any of the WAIS subtests; Block Design, Vocabulary, and Arithmetic were next in order, with *g* loadings of +0.69, +0.64, and +0.64, respectively (P. A. Vernon 1983).

The robustness of *g* across diverse test batteries was shown long ago in a study by Garrett, Bryan & Perl (1935) who factor-analyzed a battery of six varied memory tests (meaningful prose, paired-associates, free recall of words, digit span, memory for forms, and memory for objects) and extracted the general factor. This battery of tests was then factor-analyzed along with four other diverse tests not especially involving memory (motor speed, vocabulary, arithmetic, and form board). The *g* loadings of the memory tests in the two analyses correlated .80. The overall correlation between *g* factor scores based on just the memory tests and *g* factor scores based on just the nonmemory tests was .87. This is evidence that the *g* of the six memory tests is very close to the *g* of the nonmemory tests. To be sure, the memory tests were not as highly loaded on *g* (average *g* loading = .42) as the vocabulary and arithmetic tests (average *g* loading = .65), but what little *g* the memory tests have is much the same *g* as found in the nonmemory tests. One would like to see larger-scale studies of this type based on many diverse psychometric tests to determine the variance of correlations between *g* factor scores extracted from different nonoverlapping sets of tests, controlling for reliability.

A set of data provided by R. T. Osborne (personal communication) but not used in the target article, since it is unpublished data, lends support to **Cattel**'s conjecture that, when $g_c$ and $g_f$ can be clearly distinguished by including in the factor analysis a large enough number of the types of tests that will permit the emergence of these two factors, the tests' loadings on $g_f$ would be more highly correlated with the black–white differences than the loadings on $g_c$. Osborne's battery included seven of the most "fluid" tests from the Educational Testing Service's "Kit of Reference Tests for Cognitive Factors" (French, Ekstrom & Price 1963) (Cube Comparisons, Identical Pictures, Formboard, Surface Development, Spatial, Paper Folding, and Object Aperture). The "crystalized" tests in the battery were the Calendar Test, Arithmetic, the Wide Range and Heim Vocabulary Tests, and Spelling. All 12 tests were given to 608 white and 246 black urban school children. Factor analyses with varimax rotation, performed separately in each group, yielded two orthogonal (i.e., uncorrelated) factors clearly identifiable as $g_f$ and $g_c$, both of which showed high congruence between the black and the white samples. The Spearman hypothesis was examined separately for $g_c$ and $g_f$. The correlation between tests' $g_c$ loadings and the mean black–white differences is −0.24 for white $g_c$ loadings and −0.02 for the black; neither *r* is significant. The correlation between loadings on $g_f$ and the black–white difference is +0.56 ($p < .05$) for whites and +0.42 ($p < .10$) for blacks. Thus, the mean black–white differences on these 12 tests are more highly related to the tests' loadings on $g_f$ than on $g_c$. This result seems to contradict the popular belief that the black–white difference on tests largely involves differences in scholastic learning as characterized by the "crystalized" component of variance in test scores. There is some ambiguity in this study, however, owing to the fact that virtually all the nominal $g_f$ tests are also known to involve spatial visualization ability ($g_v$) as well as $g_f$; since nonspatial fluid tests were not included, $g_f$ and $g_v$ could not be distinguished, and so what appears as $g_f$ is actually some amalgam of $g_f$ and $g_v$. How closely the black–white difference is associated with each of these components separately is not known.

Another study (Jensen 1973b) of large representative samples totaling about 200 white, black, and Mexican-American Californian school children used 17 tests which included nonspatial as well as spatial tests of "fluid" ability (Lorge-Thorndike Nonverbal IQ, Raven Matrices, Figure Copying), three short-term memory tests, and typical "crystalized" tests (Lorge-Thorndike Verbal IQ and the Stanford Achievement battery of seven scholastic achievement tests). A number of socioeconomic indices (Gough Home Index) were also included. Varimax factor analysis yielded four orthogonal factors corresponding to $g_c$ (Verbal IQ and Achievement Tests) and $g_f$ (nonverbal tests), as well as a rote memory factor and a socioeconomic status factor. The mean factor scores of each of the populations on each of the factors are shown in Figure 1. The black–white difference in mean factor scores scarcely differs between the $g_c$ factor (verbal IQ and achievement) and the $g_f$ factor (nonverbal IQ). It should be noted that these are uncorrelated factors. This and other evidence, I believe, drastically undermines **Gustafsson**'s criticism that the differing compositions with respect to $g_c$ and $g_f$ of the various test batteries used to test Spearman's hypothesis has resulted in the hypothesis's remaining largely untested.

**Jones** cites an article (Jones 1984), which I have not yet seen, showing that "the average scores of the nation's black students on aptitude and achievement tests have steadily risen, relative to average scores for white students, over the past 15 years." The basis for this claim will have to be reconciled somehow with the recently announced results of the Armed Services Vocational Aptitude Battery (ASVAB), a set of ten aptitude and achievement tests administered to a large national probability sample representative of American youths ages 16 to 23 years (Office of the Assistant Secretary of Defense 1982). The mean black–white differences (in standard score units) on some of the ASVAB scholastic achievement tests are Arithmetic Reasoning 1.16, Word Knowledge 1.30, Paragraph Comprehension 1.08, and General Science 1.23. These differences are at least as large as the black–white difference on the Army Alpha at the time of World War I or on the Army General Classification Test in World War II. If there is a genuine discrepancy between

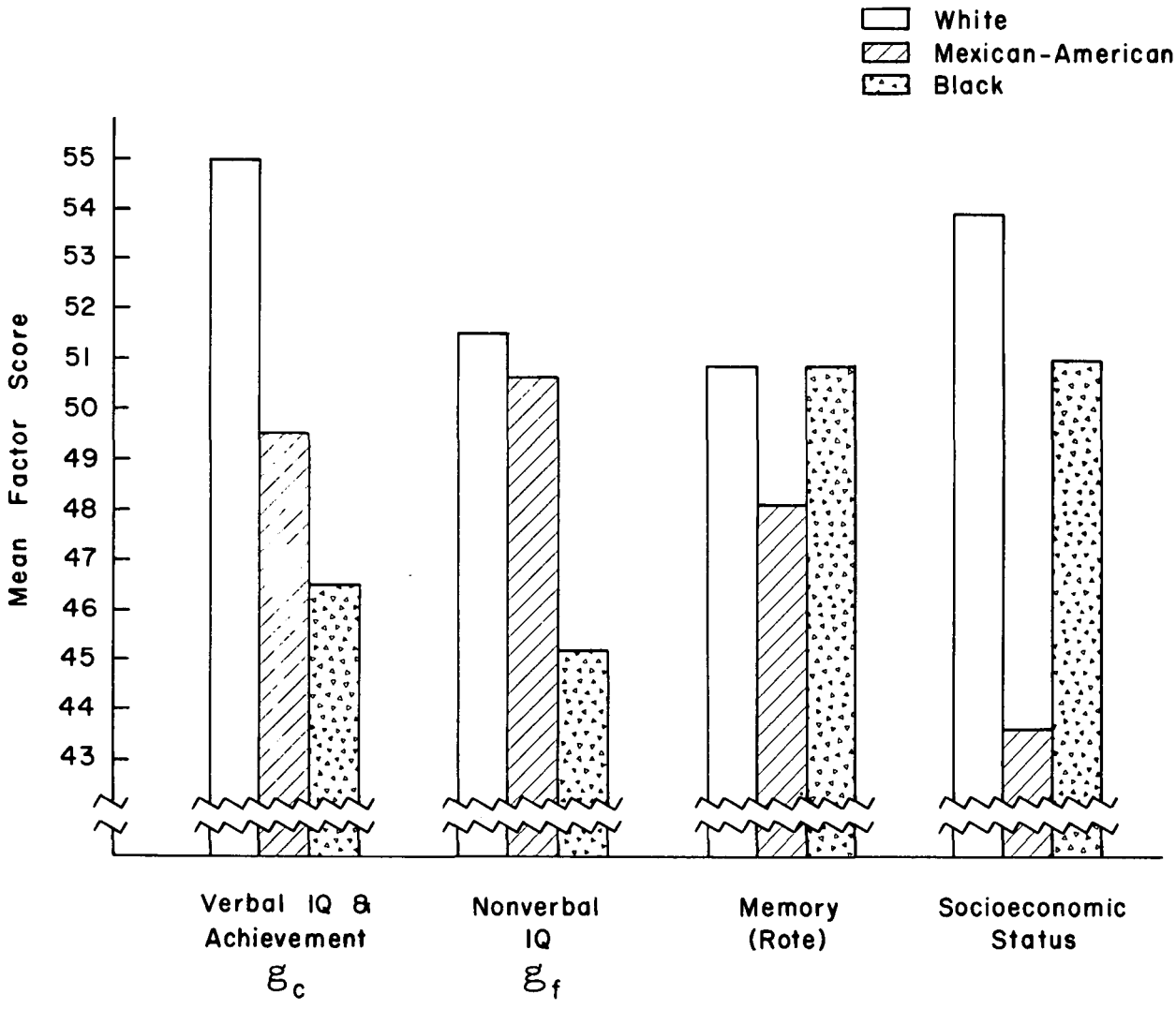


Figure 1. (Response). Mean factor scores (mean = 50, $\sigma$ = 10 within each grade level) for four variables, comparing white, black, and Mexican-American samples in grades 4, 5, and 6. The factor scores are orthogonal; that is, the scores on any one factor reveal differences between subjects who are statistically equated on the three other factors. (From Jensen 1971, Table 6.)

Jones's test results and the recent ASVAB test results, the discrepancy may be at least partly explainable in terms of Spearman's hypothesis; that is, the ASVAB tests may be more highly g-loaded than Jones's tests.

**Borkowski & Maxwell** claim that although a relationship between tests' *g* loadings and the size of black–white differences has been demonstrated, it has not been shown that the black–white difference is predominantly a difference in *g*, and hence the "weak" form of Spearman's hypothesis remains untested. They have apparently overlooked the study by Jensen and Reynolds (1982) that explicitly apportions the total between-group (black–white) variance to each of the orthogonalized hierarchical factors that emerged from a Schmid-Leiman factor analysis of the WISC-R. This study, based on the national standardization sample of the WISC-R (1868 whites and 305 blacks), showed that the black and white groups differed significantly in mean factor scores on all four of the common factors extracted from the WISC-R: *g*, verbal, performance, and memory. But in terms of the total variance between groups accounted for, the *g* factor accounted for more than seven times as much intergroup variance as the other three factors combined. The four common factors together contribute 89% of the total intergroup variance; the remaining 11% is due to the specificity of the 13 subtests. The same kind of analysis, which was based on factor scores for every subject, was impossible in the ten other studies, for which the scores of individuals were not available. The weak form of Spearman's hypothesis, however, could be further investigated in these studies by including in the test intercorrelation matrix the point-biserial correlations of the black–white dichotomy with each of the tests and then factor-analyzing the matrix to see precisely the magnitudes of the loadings of the black–white variable on each of the orthogonal factors extracted from the matrix. When this analysis is done with the WISC-R data, the results, of course, are completely consistent with those I have just reported, showing the black–white variable to have by far the largest loading on *g*. It is hard to imagine that very different outcomes would be found in the ten other test batteries, but in order to leave no doubts about the answer to this question, I will do the required factor analyses and report the results in Continuing Commentary.

**Animal intelligence.** It is difficult to evaluate **Macphail**'s claim that there is nothing resembling *g*, or individual

differences in intelligence, either between or within different species of nonhuman vertebrates. Any behavioral differences that might be interpreted as differences in cognitive ability or in some general capacity for dealing with complexity, it seems, can also be attributed to species differences in specific sensory and motor capacities or to differing instincts and drives. The literature on comparative psychology, I believe, leaves much room for doubting Macphail's claim, although the null hypothesis, which Macphail seems to favor, may be difficult to reject definitively at present. The main problem is one of devising tests that are deemed equally appropriate across species which differ widely in sensory and motor equipment and in appetites and instinctual behaviors. The problem will have to be debated and resolved empirically, if possible, by experimental comparative psychologists and ethologists. The speed of acquisition of learning sets has been found to be related to intelligence in humans (Hunt 1961, p. 83) and also shows clear inter- and intraspecies differences. As Harlow (1959) has observed, "All existent LS [learning set] data on all measured species are in keeping with the anatomical data bearing on cortical complexity, and it is obvious that LS techniques are powerful measures for the intellectual ordering of primate and possibly even nonprimate forms" (p. 507). Interspecies differences in complexity of behavioral capacities are related to brain size (in relation to body size) and to the proportion of the brain not involved in vegetative or autonomic and sensorimotor functions. According to Jerison (1973), development of the cerebral cortex, the association areas, and the frontal lobes parallels species differences in behavioral complexity. It has been found that the tests which have shown differences in problem-solving capability between monkeys and apes, and even individual differences between chimpanzees, have shown the same rank order of difficulty when they are given to human children as when they are given to apes; this suggests that the tests involve similar capacities across species (Viaud 1960, pp. 44–45).

**Macphail** harks back to Spearman's (1923, p. 346) original notion of *g* as a kind of "mental energy." Although Spearman intended this description merely as an analogy or metaphor, the notion still has intuitive appeal. High-*g* persons actually give the appearance of possessing more spontaneous mental energy, which they bring to bear on almost everything they do of a cognitive nature, and they also seem to be more persistently active in cognitive ways. But these characteristics may only be the by-products of their greater speed and efficiency of information processing. Equating *g* with drive, formulated as Hull's "big *D*," as suggested by Macphail, would seem to run into difficulty with the Yerkes-Dodson law (Yerkes & Dodson 1908), which is the now well-established empirical generalization that the optimal level of drive (*D*) for learning or performance of a task is inversely related to the degree of complexity of the task; that is, a *lower* level of *D* is more advantageous for the performance of more complex tasks. In this respect, *D* is just the opposite of *g*. The *g* loading of tasks increases with task complexity, and persons who score highest in the most *g*-loaded tests are more successful in dealing with complexity. From what research has taught us about Hull's *D* and the Yerkes-Dodson law, one would not predict high-*D* persons to perform like high-*g* persons as a function of task complexity. In humans, changes in drive and arousal are reflected in pupillary dilation. Ahern and Beatty (1979) measured the degree of pupillary dilation as an indicator of effort and autonomic arousal when subjects are presented with test problems. They found that (1) pupillary dilation is directly related to level of problem difficulty (as indexed both by the objective complexity of the problem and the percentage of subjects giving the correct answer) and (2) subjects with higher psychometrically measured intelligence show less pupillary dilation to problems at any given level of difficulty. (All subjects were university students.) Ahern and Beatty concluded:

> These results help to clarify the biological basis of psychometrically-defined intelligence. They suggest that more intelligent individuals do not solve a tractable cognitive problem by bringing increased activation, "mental energy" or "mental effort" to bear. On the contrary, these individuals show less task-induced activation in solving a problem of a given level of difficulty. This suggests that individuals differing in intelligence must also differ in the efficiency of those brain processes which mediate the particular cognitive task. (P. 1292)

**Unitarianism versus componentialism.** Questions are raised by both **Brand** and **Nichols** concerning whether *g* variation has unitary or multiple causation, and to what extent it arises from polygenic effects or from correlated environmental influences. These questions are also implicit in several other commentaries. They are really the crux of current theorizing about *g*. These issues are simply unresolved at present, but progress is being made. I do not see a sufficient empirical basis as yet for predicting whether the physiological substrate of *g* will eventually turn out to be some "unitary" feature of neural activity (e.g., cortical conductivity, speed of synaptic transmission, number of neurons, amount of branching, number or organization or complexity of cell assemblies, or capillary blood supply to the cortex) or the resultant of many such features. The well-established fact of the genetic heritability of *g*, however, makes it virtually certain that some substantial proportion of the *g* variance must ultimately find explanation at the neurophysiological level. Cognitive componential theory in all its contemporary forms represents a different level of analysis; it is a behavioral analysis of various cognitive tasks in terms of a limited number of abstracted information processes, or "components," having the status of intervening variables or psychological constructs that are hypothesized to mediate or execute different cognitive tasks. These hypothesized components, or information processes, are operationally definable, and individual differences in them are measurable, at least indirectly, by means of various chronometric techniques. The *g* yielded by factor analysis of psychometric tests, according to the componential view, results from there being certain elementary cognitive processes (and perhaps also metaprocesses) that are required for successful performance on virtually all test items. But measures of the elementary cognitive tasks are themselves intercorrelated, and when factor analyzed they yield a *g* that is correlated with the *g* of psychometric tests. Hence there is a kind of infinite regress of task intercorrelations getting at essentially one and the same *g*, at times more or less obscured or attentuated by task

specificity and measurement error. At the very end of this regress of *g* across levels of analysis, presumably, is some physiological substrate, the nature of which is still highly speculative. But we will probably not have a scientifically satisfying explanation of *g* until *g* has been clearly linked to its biological structures or physiological mechanisms. This field is wide open for theoretical speculation and empirical investigation. I do not rule out the possibility, favored by **Brand**, that the basis of *g* at this level could be something much simpler than what we can observe at the psychological or behavioral level of analysis, just as the basic cause of a disease is often much simpler than its multifarious symptoms.

**Evoked potentials and *g*.** One cannot deny **Callaway**'s assertion that brain electrical potentials, or evoked potentials, are not *necessarily* correlated with intelligence. **Carlson** expresses similar caution at this stage of this research. It is one of the primary aims of current research in this field to discover the specific procedural conditions that will yield the most substantial correlations between certain aspects of the average evoked potential (EP) and psychometric *g*. A recent study by Haier, Robinson, Braden & Williams (1983), for example, has identified various experimental conditions and methods of measurement that have resulted in some of the inconsistent findings in this field. Haier et al. identify those particular conditions that show the highest correlations between EP and IQ. They conclude:

> Perhaps, the most startling conclusion suggested by this body of work is not just that there is a relationship between brain potentials and intelligence, but that the relationship is quite strong. This supports the proposition that the variance of intelligence, with all its complex manifestations, may result primarily from relatively simple differences in fundamental properties of central brain processes. (P. 598)

**Schafer**'s comment provides further striking evidence of the relation between certain parameters of the EP and psychometric *g*. Not only do his data show an overall multiple correlation of +0.64 (or +0.80 corrected for restriction of IQ range in his sample) between the EP parameters and the WAIS Full Scale IQ, but more importantly they also show that the degree to which each of the 11 subtests loads on the *g* factor is directly related to the degree of each subtest's correlation with the EP. Figure 2 shows this relation for the EP habituation index, as defined by Schafer. (The *g* factor here is estimated by the first principal component, provided by Schafer.) Correcting the correlation for attenuation with the reliabilities of the WAIS subtests in the standardization sample results in a lowering of the correlation in Figure 2 from +0.897 to +0.891. Partialing out the subtest reliabilities produces exactly the same result for these data. Moreover, this is not an isolated finding. Eysenck and Barrett (1985), measuring a different parameter of the EP, reported a correlation (Spearman's rho) of +0.95 between WAIS subtests' *g* loadings and the subtests' correlations with the EP measure. It is probably more than sheer coincidence that the correlation between Schafer's EP habituation index and the WAIS subtests shows a rank-order correlation of +0.59 ($p < .05$) with the degree of inbreeding depression (a purely genetic effect) found on the homologous subtests of the WISC (Jensen

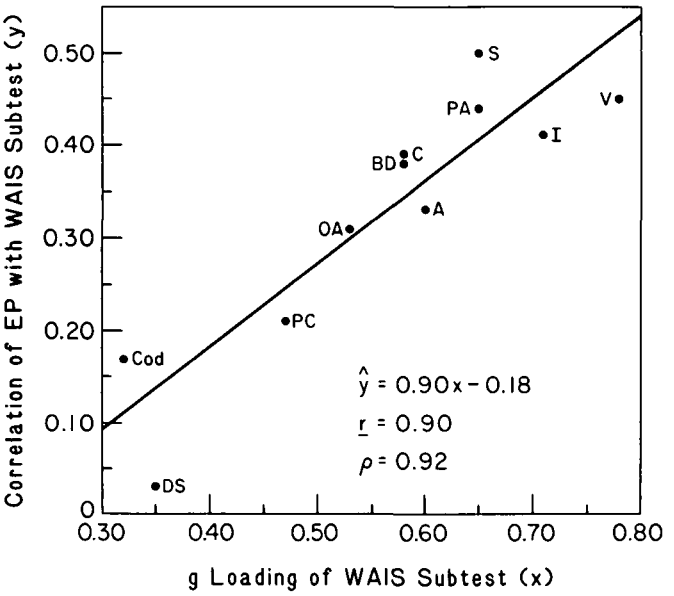


Figure 2. (Response). Correlation of the habituation index of the evoked potential (EP) with Wechsler Adult Intelligence Scale (WAIS) subtests plotted as a function of the subtests' *g* loadings, in Schafer's study. WAIS subtests: 1 – Information (I), 2 – Comprehension (C), 3 – Arithmetic (A), 4 – Similarities (S), 5 – Digit Span (DS), 6 – Vocabulary (V), 7 – Coding (Cod), 8 – Picture Completion (PC), 9 – Block Design (BD), 10 –Picture Arrangement (PA), 11 – Object Assembly (OA).

1983a). We can eagerly look forward to the working out of **Callaway**'s promising suggestions concerning the use of "psychopharmacological tools" for manipulating the biological variables underlying information processes. This biological-analytical approach is a promising avenue toward understanding the physiological substrate of *g*.

## Chronometric correlates of *g*

In connection with the evoked potential studies just mentioned, it is worth noting a parallel phenomenon based on the correlation of reaction time (RT) with Wechsler subtests. P. A. Vernon (1983) extracted the general factor from a battery of elementary cognitive tasks (ECTs) in which RT was the dependent variable. The ECTs were so simple that the largest mean RTs were less than one second. The ECT general factor was substantially correlated with the WAIS Full Scale IQ, and the correlation of the general speed factor with the various WAIS subtests was related to the subtests' *g* loadings. Especially interesting is the fact that no other factors of the WAIS besides *g* showed any correlation with the ECT general speed factor. Since the target article was written, a similar recent study has come to my attention, based on the WISC-R in a sample of 59 elementary school pupils (Hemmelgarn & Kehle 1984). An apparatus very similar to that shown in Figure 8 of the target article was used. Individual differences in the slope of RT as a function of bits of information, interpreted as a measure of rate of information processing, were correlated with each of the WISC-R subtest scores (with chronological age partialed out). This profile of 12 correlations (i.e., subtests and slope of RT) showed a correlation of −0.80 ($p < .05$) with the profile of subtests' *g* loadings. The overall correlation between RT slope and WISC-R Full Scale IQ was only

$-0.32$ ($p < .01$); but a much higher correlation than this could hardly be expected, because it has been generally found that the slope parameter has the lowest reliability of any of the individual difference measures derived from this RT paradigm. (See Jensen, 1982a, 1982b, for detailed discussions of this RT paradigm.) Most probably, low reliability is the answer to **Carlson**'s observation that correlations between *g* and RT have not consistently shown the predicted increasing relationship across bits of information in all studies. When the means of groups differing in average IQ are used to examine slope instead of the much less reliable measures of individual differences, however, the results have been quite consistent in showing that in low-IQ groups the slope of RT across bits is greater than in high-IQ groups even when both of the contrasted groups are above the general population average in IQ.

**Strategy of RT studies.** There is criticism from **Carr & McDonald, Posner,** and **Rabbitt** of the fact that my presentation of correlations between RT measures and various elementary cognitive tasks (ECTs) and psychometric scores has not emphasized the same kind of analytic technique (consisting mostly of variations of Donders's subtraction method) commonly used in experimental mental chronometry. This approach is nicely summarized by Carr & McDonald. Hypothetical cognitive processes are measured indirectly by subtracting the RT for a task in which a particular process is believed to be absent from the RT for a task in which the process is believed to be required for successful performance. The remainder is a measure (usually in msec) of the time taken by the hypothesized mental process on which the two tasks are presumed to differ. I agree that this methodology is highly desirable and ultimately essential in the chronometric study of individual differences and their relation to psychometric variables. However, I considered it a highly inefficient strategy for initially exploring relationships between chronometric and psychometric variables. Those investigators who have pursued only the experimental psychology of RT, divorced from its possible relationship to individual differences in psychometric factors, may have forgotten that just a few years ago it was conventional wisdom in psychology that RT had no relationship to intelligence. Almost every psychology undergraduate has been taught in lectures and textbooks that the Galton-Cattell (i.e., James McKeen Cattell) "brass instrument" attempt to measure intelligence by means of RT and various tests of sensory discrimination was an utter failure, without learning specifically why it was a failure, and that only very complex or achievement-type tests are capable of reflecting (or defining) what psychologists mean by "intelligence." This has now been conclusively disproved by a great many recent studies. But prior to about ten years ago, I found surprising resistance to – and often scoffing rejection of –the idea that Galton and Cattell may have been right, or at least partly right, after all. It was apparent that a correlation between RT and psychometric *g* would take a lot of "proving" even for most psychologists to come to agree that there might be something worth investigating in this realm. A broad-gauged or "shotgun" search for correlations and mean differences between criterion groups selected from different sectors of the IQ distribution seemed the best strategy. Why invest a great deal of experimental refinement in some chronometric technique before establishing that at least some of the RT parameters it yields are significantly correlated with the individual difference variable of primary interest, that is, psychometric *g*, with all its obviously important scholastic, occupational, and social correlates? Whatever correlations might exist would be revealed by the raw RT measures (and such simple parameters as slope and intraindividual [trial-to-trial] variability in RT) just as well as, if not better than, the complex derived measurements of the processes hypothesized to be involved in performance on the chronometric tasks. These complex measures usually consist merely of different linear combinations of the raw RT measurements, and so any correlation that the derived measures might have with test scores would also necessarily be revealed by multiple regression analysis of the raw RT measurements. Moreover, correlational studies require good-sized samples, which, at least in exploratory research, necessitates using relatively few RT trials per subject, at the expense of achieving high reliability of individual measurements. Derived measures, being based largely on difference scores, magnify the effects of unreliability and hence further attenuate the possible correlations between RT and psychometric variables, rendering the search for correlations liable to Type II error. It is surprising that **Nettelbeck** does not seem to have noticed how seriously this very kind of Type II error has vitiated the results of the recent study by Borkowski and Krause (1983), which Nettelbeck views so uncritically. I have noted the shortcomings of this study in detail elsewhere (Jensen 1985).

Another factor in my reluctance to dive into a componential type of analysis of chronometric data in this initial exploratory stage of our research is based on what I have learned from R. J. **Sternberg**'s experience. This is the fact that there is a general RT factor (or "regression constant," as Sternberg usually terms it) in a variety of chronometric variables that is more highly correlated with psychometric *g* than most of the measurements representing specific cognitive processes (or "components," in Sternberg's terminology). In summarizing the research on the componential analysis of chronometric tasks and the correlation of components with IQ, or *g*, Sternberg and Gardner (1982) make the following observation:

> A result that at first glance appears most peculiar has emerged from many of these task analyses. . . . The regression intercept, or global "constant," often turns out to be as highly correlated or more highly correlated with scores from IQ tests than are the analyzed parameters representing separated sources of variance. Since the constant includes speed of response, e.g., button pressing, one could interpret such results trivially as indicating that motor speed is an essential ingredient of intelligence. A more plausible interpretation, and, as it will turn out, one more consistent with the bulk of the data, is that there are certain constancies in information-processing tasks that tend to be shared across wide variations in item types. We suggest that the search for the general component(s) and the search for the general factor are one and the same search–that whatever it is that leads to a unitary source of individual differences across subjects also leads to a unitary source of difference across stimulus types. (Pp. 232–33)

So before focusing on specific cognitive processes, or components, we have tried to establish firmly the correlation between the general factor of RT tasks and psychometric *g*. We are interested in whatever significant correlations we find, regardless of whether or not they are consistent with any theoretical preconceptions that we or anyone else may have had. When critics gleefully point out some theoretically unexpected effect, such as that movement time (MT) is sometimes about as highly correlated with *g* as RT, or that the RT intercept shows a higher correlation with *g* than does slope in some samples, as if they had scored a crucial point, I cannot keep from smiling. Are such findings to be put down as a loss? Theories are so tentative in this field at present that one must place more emphasis on discovering empirical relationships than on testing any specific theory. I regard any significant and replicable correlations that are unexpected in terms of general theoretical preconceptions as no less interesting than those that confirm a particular theoretical preconception. We have indeed had many surprises in our RT research so far; when they are reliable and replicable they are perfectly suitable material for theory and further inquiry. A certain "critical mass" of firmly established empirical relationships seems to be a necessary prerequisite for efficiently pursuing the kind of theory-oriented strong-inference research extolled by **Callaway,** which I agree is called for in the next phase of this program of research, now that it has been quite thoroughly demonstrated that our several chronometric paradigms yield various individual difference parameters that are indeed reliably related to psychometric *g*.

**Specific criticisms of the RT research.** It is always possible for critics to ignore the overall consistencies in a number of related studies and to invent ad hoc hypotheses that would seem to explain, or more usually to explain away, the results of any particular study. I am not willing to agree, however, that, because it is theoretically impossible to construct an ideally perfect lens, or because there is always some degree of atmospheric perturbation of light rays, astronomy is an altogether impossible science. The fact that it may be possible to find certain experimental paradigms, conditions, or testing procedures under which chronometric variables are not significantly correlated with psychometric variables is of no great concern, since we are seeking those conditions which *do* show correlations. And we are finding them. From our standpoint, those RT conditions which fail to yield correlations with *g* are of interest for that reason alone, but they have no theoretical refutational power whatsoever, as long as other conditions do in fact show reliable, replicable correlations with *g*.

**Rabbitt** surmises that the experimental separation of RT and MT in our chronometric procedures could result in a strategy, presumably adopted by the more intelligent subjects, in which there is a trade-off between RT and MT, such that subjects can shorten their RTs by responding *before* actually making a choice decision and then "hovering" to make the decision before executing the MT part of the response. **Carr & McDonald** raise essentially the same question. If this strategy were indeed in effect, we should predict a *negative* correlation between RT and MT both *within* subjects (from trial to trial) and *between* subjects (i.e., the subjects with faster RTs showing slower MTs), as well as correlations of opposite sign between *g* and RT and MT. We have long since examined all of these possibilities in our data and the results do not bear them out in the least: RT and MT are completely uncorrelated *within* subjects and *positively* correlated *between* subjects; and we have never found correlations of RT and MT with intelligence that are of opposite sign. Also of considerable interest is our finding that variation in task complexity is strongly reflected in RT but hardly at all in MT. A recent study in our laboratory, involving 14 variations in task complexity (all yielding median RTs within the range of about 600 to 1300 msec), found that the RTs on each of the tasks were much more highly correlated with Raven Matrices scores than with the MTs on the same tasks (Paul 1985). Rabbitt also conjectures that group differences in choice RT might diminish or disappear if RT trials were continued long enough for the groups to reach asymptotic levels of RT. In one study (see Jensen 1982b, p. 105) in which a group of 10 subjects was run on the Hick choice RT paradigm for a total of 540 trials spread over 9 practice sessions, there was no significant change in mean RT beyond the first session, which was the same as our standard testing procedure. We have not yet examined the effects of extended practice on the other RT tasks in the battery. The asymptotic study that Rabbitt recommends was actually done by Noble (1969), who measured RTs on 106 black and 106 white age-matched school children given 160 trials on a four-choice discrimination RT task. The groups differed significantly (whites faster), without the least indication of asymptotic convergence of the groups' mean RTs, as shown in Figure 3.

The study by Vernon and Jensen (1984) could not, of course, be reported in every detail in the target article, but the variances (or *SD*s) and correlations of the various tasks and other information that **Rabbitt** regards as important are provided in the original article. Both **Rabbitt** and **Posner** note that tasks SD2 (physically same–different words) and SA2 (synonyms–antonyms) involve verbal content, and they claim that the verbal content, rather than the tasks' intrinsic information-processing difficulty, is probably responsible for the black–white difference on these tasks. The ambiguity in interpreting this result is fully recognized by Vernon and Jensen (1984, p. 421). Other studies designed to resolve this ambiguity are already in progress. It will be surprising to me if Posner's conjecture that differential reading skill of blacks and whites, independent of *g*, would account for the black–white difference on tasks SD2 and SA2. One statistical test would be to regress out that part of the variance in reading skill which is independent of *g* (assessed by nonverbal tests) from the RT variables and see whether a significant black–white difference in mean RTs remains. Other research indicates that when *g* is regressed out of scores on verbal tests, the black–white difference virtually disappears. That is, the difference in reading skill seems largely to reflect the more general black–white difference in *g*.

Experimental chronometricians (**Nettelbeck, Poortinga, Posner, Rabbitt**) are concerned with the phenomenon known as "speed–accuracy trade-off," suggesting that perhaps the brighter subjects adopt a strategy of sacrificing accuracy for speed, thereby showing faster RT and a higher error rate. But this trade-off seems to be mainly a *within*-subjects phenomenon, accounting for

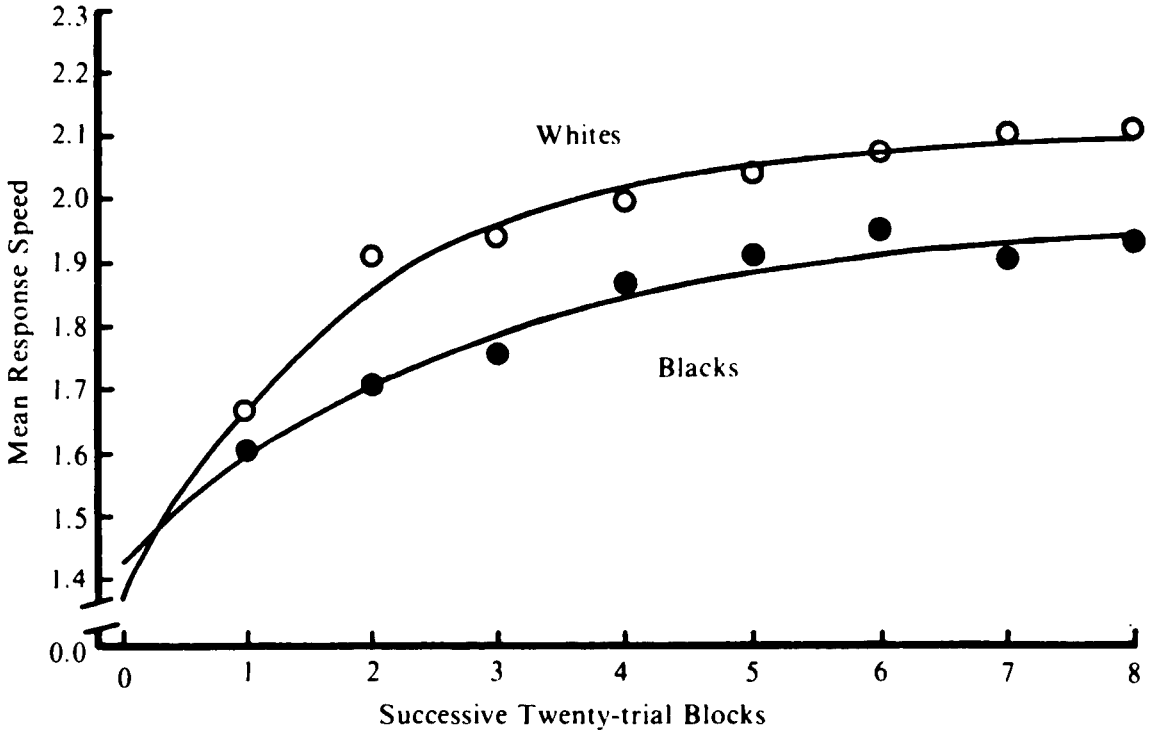


Figure 3. (Response). Mean response speed (reciprocal of RT) in successive 20-trial blocks on a 4-choice RT test. Each curve based on 106 children. (From Noble 1969.)

*negative* correlations (*within* subjects) between RTs and error rates under different levels of task difficulty. It has not been a problem at all in the interpretation of the correlation between individual differences in RT and *g*, because the *between*-subjects correlation of RT and error rate is a *positive* correlation, and both RT and error rate are negatively correlated with *g*. That is, the brighter subjects are both faster and more accurate than the less bright subjects; we have never found any evidence of a speed–accuracy trade-off *between* subjects in our analyses of RT data. These relationships can perhaps be seen more clearly as depicted in Figure 4. On the *simple task*, hypothetical persons A, B, and C are shown to have the same short RT and low error rate. On the *complex task*, the latent ability differences between A, B, and C are manifested as variation in their RTs and error rates. Their performances, as reflected jointly by RT and errors, will tend to fall somewhere on each of the arcs that describe the speed–accuracy trade-off and are different for each person. If the same low error rate of the simple task is to be maintained for the complex task, the RT is greatly increased for all persons (vertical line = zero speed–accuracy trade-off). If the RT in the simple task is to be maintained in the complex task, the error rate is greatly increased for all persons (horizontal line = 100% speed–accuracy trade-off). So the arc for each person describes an *inverse* relationship (or *negative* correlation) between RT and error rate. But *between* persons, RT and error rate show a *direct* relationship (or *positive* correlation). The line marked *X* in Figure 4 indicates a fairly high speed–accuracy trade-off for a typical RT study, if the error rate (on the abscissa) is assumed to range between *zero* and *chance*. Thus the shaded area represents the most desirable region for performance when studying individual differences in RT in that it spreads out individual differences in RT much more than in error rate, a feature observed in all of our RT studies. Hence the observed correlation between RT variables and *g* can in no way be accounted for in terms of speed–accuracy trade-off.

**Jones** complains that the figures showing mean group differences on the various chronometric tasks express the differences directly in terms of milliseconds, rather than in standard deviation (σ) units. I had used the raw RT

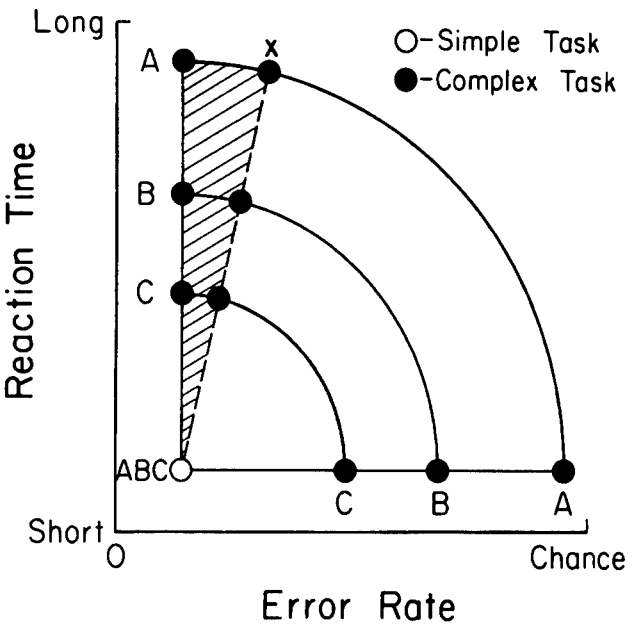


Figure 4. (Response). The idealized relationship between RT and error rate for simple and complex tasks. The arcs describe the speed–accuracy trade-off for hypothetical persons A, B, and C, who are shown here as performing equally on the simple task. Shaded area represents most desirable region of speed–accuracy trade-off for RT studies.

differences to take advantage of a luxury that is generally denied for ordinary psychometric tests, namely, a true ratio scale, which RT represents, so that the mean group differences in RT are differences in real time units, with equal intervals and a true zero point. The results depicted in Figures 11 and 12, it turns out, remain essentially the same when differences are expressed in $\sigma$ units. In Figure 12, for example, when the group differences (vocational college versus university) on the tasks, expressed in $\sigma$ units, are plotted as a function of task complexity as indexed by mean RT, the Pearson correlation is +0.92 ($\rho$ = +0.93), as compared with +0.97 when the RT differences are expressed in msec. If instead of differences we use the ratio of vocational college/university RTs, the correlation is +0.89 ($\rho$ = +0.95); and if the RTs are subjected to a logarithmic transformation (which tends to make the standard deviations and means uncorrelated), the corresponding correlation becomes +0.93 ($\rho$ = +0.95). In other words, no matter what the scale is on which the group differences are expressed, the group differences are found to increase as a function of task difficulty or complexity. (The same thing is true of Figure 11.) [I am grateful for Jones's noting the errors in the target article's Figure 10, which have been duly corrected in the published version.] The other questions raised by Jones about this study are answered in the original Vernon and Jensen (1984) article.

**Poortinga** believes that cultural factors may affect RT in ECTs and that such tasks as simple RT may be culturally biased and hence "nonequivalent" across different populations. But the lack of evidence for cultural bias with respect to the American black and white populations in much more complex and culture-loaded psychometric tests makes it an improbable hypothesis that cultural bias would be significantly implicated in ECTs. Cultural bias could be investigated by much the same methods as have been applied to conventional tests (Jensen 1980a). Poortinga infers bias on the basis of theoretical preconceptions of the pattern of group differences one should expect for various RT parameters. This puts too much faith in the present theories of RT and ECTs. For the time being, I would avoid theoretical preconceptions about which parameters should be most meaningful and take a more direct empirical approach. This would consist of looking at differences in RT parameters between different population samples that are hypothesized to differ culturally in ways that affect performance in ECTs and comparing the pattern of differences with the corresponding patterns found in pairs of groups that are selected to be high and low in psychometric *g* but are culturally equivalent. Ideally, one could use groups of full siblings reared together, with one member of each sib pair assigned to the low-*g* group and the other member assigned to the high-*g* group. These two comparison groups would be as culturally equivalent as possible. If the two supposedly culturally different population samples show essentially the same pattern of RT differences on a number of ECTs as the culturally equivalent groups that were selected to differ in *g*, then we would be forced either to reject the cultural bias hypothesis *or* to hypothesize that the cultural difference perfectly mimics the *g* difference between two culturally equivalent groups. With enough different ECTs, the latter hypothesis becomes highly implausible. I would like to see this type of study performed with the set of RT tasks that were used in Poortinga's (1971) own interesting study.

**RT and athletic skill.** The black–white differences in response latencies on some of the elementary cognitive tasks is called into question by **Das** and **Whimbey** on the ground that a relatively large proportion of topnotch athletes and Olympic gold medalists are black. First, it is a mistake to try to explain a given phenomenon (black–white differences in RT) in terms of another even more complex and less well understood phenomenon (athletic skill). And a phenomenon observed in one realm (the athletic field) certainly cannot refute a questionably related phenomenon observed in another realm (the psychological laboratory). Second, the exceptional Olympic-level athletes are highly selected from their respective populations, and their particular talents may represent other features of the population distribution of ability than the central tendency, such as the variance, which would affect the remote tails of the distribution from which exceptionally talented individuals are selected. Third, the argument presumes that the order of RTs (in the range of about 200 to 1200 msec) represented in our studies constitutes a sizable proportion of the variance in athletic skills. This is most unlikely. RT evidently has much more to do with *g* than with athletic prowess. Noble (1978) lists a large number of physical fitness and body build factors, independent of psychomotor and perceptual factors, that are involved in varying degrees in different athletic skills, which generally require sequential integration of numerous separate movements of large muscle groups, whole-body coordination, and the like. It may seem even more surprising to Das and Whimbey that blacks have been found to perform significantly less well than whites even on the pursuit rotor, a simple motor learning task (Noble 1978, pp. 346–47; Payne & Turkat 1982). Apparently, very fast RT is not necessary for becoming the greatest boxer of all time. According to Keele (1973, as cited by Hunt 1976, p. 238), "Muhammad Ali, a heavyweight boxer who, in his prime, was lauded for his 'cat-like reflexes,' had a quite average motor reaction time."

## The genetic heritability issue

Several commentators (**Bardis, Cattell, Johnson & Nagoshi,** and **Stanovich**) bring up the genetic question. However, I have consistently treated Spearman's hypothesis as a *phenotypic* phenomenon. Strictly speaking, neither the data nor the methodology of the target article permits inferences about the relative roles of genetic and nongenetic sources of variance in the observed, or phenotypic, population differences. Stanovich is perfectly right in noting that the findings are moot regarding the *causes* of the differences. I have long since concluded that the only technically available method, at present, that would permit proper genetic inferences regarding population differences in IQ (or in any other phenotype) would be to perform a true genetic experiment, cross-mating random samples of the two populations and cross-fostering the offspring. But socially and ethically such an experiment would be wholly unfeasible and impermissible. All other feasible lines of research can at most only diminish or augment the subjective plausibility of the

hypothesis that genetic factors are involved in any particular physical or mental trait difference between populations. The broad evolutionary context of biological and behavioral variables in which **Rushton** finds remarkably systematic relationships among differences between populations of African, Asian, and European origin affords a much needed perspective for further advances in the study of human variation, although such research will unfortunately invite still more controversy and even opprobrium in the ideological climate that currently prevails in the social sciences.

Individual variation *within* populations is quite another matter, however. It is now well established that genetic factors are strongly involved in individual differences on psychometric tests. (**Bardis** is simply wrong on this issue, and he errs in believing that the estimation of heritability depends on the direct measurement of environmental factors.) But ECTs have not yet been subjected to extensive genetic analysis. The only published genetic study of ECTs that I am aware of is based on several ECTs quite similar to those described in the target article, administered to a total of 47 pairs of monozygotic and dizygotic twins reared apart, from which the authors (McGue, Bouchard, Lykken & Feuer 1984) concluded:

> The results reported here support the existence of a general speed component underlying performance on most experimental cognitive tasks which is strongly related to psychometric measures of "g," and for which there are substantial genetic effects. Although much of the relationship between psychometric test performance and processing speed may be attributed to the relationship between this general speed factor and "g," we did find evidence for a second component which loads on measures of the rate of specific cognitive processes, which was specifically associated with psychometric measures of verbal ability, and which appeared to have little or no genetic basis. (P. 256)

### The social context of *g*

The only commentator who brings Spearman's hypothesis directly and specifically into apposition with its real-life social and economic consequences, is **Cattell,** who predicts that the percentage of blacks in different occupations should be inversely related to the mean intelligence levels of persons employed in the occupations. If shown to be true, this prediction would mean, of course, that disparities in the proportional representation of black and white workers in various occupational categories are not mainly attributable to prejudice and discrimination in hiring, but are due to differences in measurable *g*-loaded abilities, whatever the cause of the differences. I have not looked into data on this point myself, but quite precise data on a range of occupations (ranging from physician and engineer to truck driver and meat cutter), directly aimed at Cattell's prediction, have been assembled by Linda Gottfredson (personal communication), a sociologist at the Johns Hopkins University. In light of Cattell's query, it would be most valuable if Gottfredson submitted this analysis to Continuing Commentary. Gottfredson's analysis, based on 1970 and 1980 statistics from the U.S. Department of Labor and the Bureau of the Census, strikingly bears out Cattell's prediction, with a near perfect rank-order correlation between the theoretically expected and the observed ratios of black to white employees in different occupations.

I suppose it is largely because of my investigating phenomena such as Spearman's hypothesis, which have such crucial and sensitive social correlates, that perhaps quite a few psychologists share in **Sternberg**'s emotional "distaste" for my study of black–white differences (also voiced in different tones by **Bardis** and **Das**). I make no apology for my choice of research topics. I think that my own nominal fields of expertise (educational and differential psychology) would be remiss if they shunned efforts to describe and understand more accurately one of the most perplexing and critical of current problems. Of all the myriad subjects being investigated in the behavioral and social sciences, it seems to me that one of the most easily justified is the black–white statistical disparity in cognitive abilities, with its far-reaching educational, economic, and social consequences. Should we not apply the tools of our science to such socially important issues as best we can? The success of such efforts will demonstrate that psychology can actually behave as a science in dealing with socially sensitive issues, rather than merely rationalize popular prejudice and social ideology.

## References


Ahern, S. & Beatty, J. (1979) Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science* 205:1289–92. [rARJ]

Allport, D. A. (1980) Patterns and actions: Cognitive mechanisms are content-specific. In: *Cognitive psychology: New directions*, ed. G. Claxton. Routledge & Kegan Paul. [THC]

Anderson, R. J. & Sisco, F. H. (1977) *Standardization of the WISC-R performance scale for deaf children.* Office of Demographic Studies Publication Series T, no. 1. Gallaudet College. [taARJ]

Ashcraft, M. H. & Stazyk, E. H. (1981) Mental addition: A test of three verification models. *Memory & Cognition* 9:185–96. [THC]

Baddeley, A. D. & Hitch, G. (1974) Working memory. In: *The psychology of learning and motivation*, vol. 8, ed. G. Bower. Academic Press. [PMAR]

Baker, J. R. (1974) *Race.* Oxford University Press. [JPR]

Bardis, P. D. (1969) The principle of instrumental parsimony. *Revue Internationale de Sociologie:* 92–101. [PDB]

Baron, J. (in press) *Rationality and intelligence.* Cambridge University Press. [JB]

Baron, J. & Treiman, R. (1980) Some problems in the study of differences in cognitive processes. *Memory and Cognition* 8:313–21. [JB]

Bengtson, V. L., Kasschau, P. L. & Ragan, P. K. (1977) The impact of social structure on aging individuals. In: *Handbook of the psychology of aging*, ed. J. E. Birren & K. W. Schaie. Van Nostrand Reinhold. [JPR]

Bethge, H. J., Carlson, J. S. & Wiedl, K. H. (1982) The effects of dynamic assessment procedures on Raven matrices performance, visual search behavior, test anxiety and test orientation. *Intelligence* 6:89–97. [JSC]

Blanchard, J. (1980) Preliminary investigation of transfer between single-word decoding ability and contextual reading comprehension by poor readers in grade six. *Perceptual and Motor Skills* 51:1271–81. [KES]

Blau, Z. (1981) *Black children/white children.* Free Press. [KES]

Bock, R. D. & Mislevy, R. J. (1981) *The profile of American youth: Data quality analysis of the Armed Services Vocational Aptitude Battery.* University of Chicago/National Opinion Research Center. [taARJ]

Borkowski, J. G. (1985) Signs of intelligence: Strategy generalization and metacognition. In: *The growth of reflection in children*, ed. S. Yussen. Academic Press. [JGB]

Borkowski, J. G. & Krause, A. (1983) Racial differences in intelligence: The importance of the executive system. *Intelligence* 1:379–95. [JGB, rARJ, TN, KES]

Borkowski, J. G. & Peck, V. A. (in press) Causes and consequences of metamemory in gifted children. In: *Conceptions of intelligence*, ed. R. Sternberg & J. Davidson. Cambridge University Press. [JGB]

Bouchard, T. J., Jr. (1982) Twins: Nurture's twice-told tale. In: *1983 Yearbook of science and the future*. Encylopaedia Britannica. [JPR]

Bouchard, T. J., Jr. & McGue, M. (1981) Familial studies of intelligence: A review. *Science* 212:1055–59. [JPR]

Braden, J. P. (1984) The factorial similarity of the WISC-R Performance Scale in deaf and hearing samples. *Personality and Individual Differences* 5:403–9. [taARJ]

Bradley, L. & Bryant, P. (1983) Categorizing sounds and learning to read–A causal connection. *Nature* 301:419–21. [KES]

Bradley, R. & Caldwell, B. (1984) The relation of infants' home environments to achievement test performance in first grade: A follow-up study. *Child Development* 55:803–9. [KES]

Brand, C. R. (1984) Personality dimensions: An overview of modern trait psychology. In: *Psychological Survey 5*, ed. J. Nicholson & H. Beloff. British Psychological Society. [CB]

(1985a) The psychological bases of political attitudes and interests. In: *Hans Eysenck: Consensus and controversy*, ed. S. Modgil & C. Modgil. Falmer. [CB]

(1985b) Intelligence and inspection time: An ontogenetic relationship? In: *The biology of human intelligence*, ed. C. Turner. Eugenics Society. [CB]

Bridgeman, B. & Buttram, J. (1975) Race differences in nonverbal analogy test performance as a function of verbal strategy training. *Journal of Educational Psychology* 67:586–90. [JSC]

Bulmer, M. G. (1970) *The biology of twinning in man*. Clarendon Press. [JPR]

Butterfield, E. C. & Ferretti, R. P. (in press) Toward a theoretical integration of cognitive hypotheses about intellectual differences among children. In: *Intelligence and cognition in special children*, ed. J. Borkowski & J. Day. Ablex. [JGB]

Callaway, E. (1975) *Brain electrical potentials and individual psychological differences*. Grune & Stratton. [EC, taARJ]

(1984) Human information processing: Some effects of methylphenidate, age and scopolamine. *Biological Psychiatry* 19:649–62. [EC]

Callaway, E., Halliday, R., Naylor, H. & Schechter, G. (in press) Effects of oral scopolamine on human stimulus evaluation. *Psychopharmacology*. [EC]

Carlson, J. S. (1983) Applications of dynamic assessment to cognitive and perceptual functioning of three ethnic groups. National Institute of Education final report, grant #NIE-G-81-0081, 100 pp. [JSC]

Carlson, J. S. & Jensen, C. M. (1982) Reaction time, movement time, and intelligence: A replication and extension. *Intelligence* 6:265–74. [JGB, taARJ, TN]

Carlson, J. S., Jensen, C. M. & Widaman, K. F. (1983) Reaction time, intelligence and attention. *Intelligence* 7:329–44. [JSC]

Carlson, J. S. & Wiedl, K. H. (1980) Applications of a dynamic testing approach in intelligence assessment: Empirical results and theoretical formulations. *Zeitschrift für Differentielle Psychologie* 1:303–18. [JSC]

Carr, T. H. (1981) Building theories of reading ability: On the relation between individual differences in cognitive skills and reading ability. *Cognition* 9:73–114. [THC]

(1984) Attention, skill, and intelligence: Some speculations on extreme individual differences in human performance. In: *Learning and cognition in the mentally retarded*, ed. P. H. Brooks, R. Sperber & C. McCauley. Erlbaum. [THC]

Carr, T. H., Brown, T. L. & Vavrus, L. G. (1985) Using component skills analysis to integrate findings about reading development. In: *New directions in child development #27: The development of reading skills*, ed. T. H. Carr. Jossey-Bass. [THC]

Carr, T. H., Pollatsek, A. & Posner, M. I. (1981) What does the visual system know about words? *Perception and Psychophysics* 29:183–90. [THC]

Carr, T. H., Posner, M. I., Pollatsek, A. & Snyder, C. R. R. (1979) Orthography and familiarity effects in word processing. *Journal of Experimental Psychology: General* 108:389–414. [MIP]

Carroll, J. B. (1980) Individual difference relations in psychometric and experimental cognitive tasks. Report no. 163, L. L. Thurstone Psychometric Laboratory, University of North Carolina. [PK]

(1981) Review of *Bias in mental testing*, A. R. Jensen. *Psychometrika* 46:227–33. [LVJ]

Cattell, R. B. (1940) A culture-fair intelligence test I. *Journal of Educational Psychology* 31:161–79. [RBC]

(1963) Theory of fluid and crystalized intelligence: A critical experiment. *Journal of Educational Psychology* 54:1–22. [RBC, J-EG, taARJ]

(1967) The theory of fluid and crystallized intelligence checked at the 5-year-old level. *British Journal of Educational Psychology* 37:209–34. [RBC]

(1971) *Abilities: Their structure, growth and action*. IPAT, Champaign, Ill. [RBC]

(1978) *The scientific use of factor analysis in behavioral and life sciences*. Plenum. [taARJ]

Chase, W. G. & Ericsson, K. A. (1981) Skilled memory. In: *Cognitive skills and their acquisition*, ed. J. R. Anderson. Erlbaum. [PMAR]

Chase, W. G. & Simon, H. A. (1973) Perception in chess. *Cognitive Psychology* 4:55–81. [THC]

Chi, M. T. H. (1976) Short-term memory limitations in children: Capacity or processing deficits? *Memory and Cognition* 4:559–72. [MIP]

Cole, M., Gay, G. & Glick, J. (1968) Some experimental studies of Kpelle quantitative behavior. *Psychonomic Monograph Supplement* 2, no. 10 (Whole no. 26):173–90. [YHP]

Coltheart, M. (1981) Disorders of reading and their implications for models of normal reading. *Visible Language* 15:245–86. [MIP]

Cooper, L. & Regan, D. T. (1982) Attention, perception and intelligence. In: *Handbook of human intelligence*, ed. R. Sternberg. Wiley. [PMAR]

Cronbach, L. J. (1957) The two disciplines of scientific psychology. *American Psychologist* 12:671–84. [CB]

(1969) Heredity, environment and educational policy. *Harvard Educational Review* 39:338–47. [JPD]

(1979) The Armed Services Vocational Aptitude Battery–A test battery in transition. *Personnel and Guidance Journal* 57:232–37. [taARJ]

Dasen, P. R. (1984) The cross-cultural study of intelligence: Piaget and the Baoulé. *International Journal of Psychology*. In press. [JSC]

Day, J. E., French, L. & Hall, L. K. (1985) Social interaction and cognitive development. In: *Metacognition, cognition, and human performance*, ed. D. L. Forrest-Pressley, G. E. MacKinnon & T. G. Waller. Academic Press. [JGB]

Dempster, F. N. (1981) Memory span: Sources of individual and developmental differences. *Psychological Bulletin* 89:63–100. [CB]

(1985) Short-term memory development in childhood and adolescence. In: *Basic processes in memory development: Progress in cognitive development research*, ed. C. J. Brainerd & M. Pressley. Springer. [CB]

Detterman, D. (1982) Does "g" exist? *Intelligence* 6:99–108. [KES]

Detterman, D. K. & Sternberg, R. J., eds. (1982) *How and how much can intelligence be increased*. Ablex. [taARJ]

Dillon, R. & Carlson, J. S. (1978) The use of activation variables in the assessment of cognitive abilities in three ethnic groups: A testing-the-limits approach. *Educational and Psychological Measurement* 38:437–43. [JSC]

Donders, F. C. (1868/69; 1969) Over de snelheid van psychische processen. Trans. W. G. Koster. In: *Attention and performance II*, ed. W. G. Koster. *Acta Psychologica* 30:412–31. [THC]

Duncan Johnson, C. (1981) P300 latency: A new metric of information processing. *Psychophysiology* 18:207–15. [EC]

Durkin, D. (1982) A study of poor black children who are successful readers. Reading Education Report no. 33. Center for the Study of Reading, University of Illinois. [KES]

Engle, R. W. & Bukstel, L. (1978) Memory processes among bridge players of different expertise. *American Journal of Psychology* 91:673–89. [THC]

Ertl, J. P. (1969) Neural efficiency and human intelligence. Final report, U.S. Office of Education Project no. 9-0105. [EC, RBC]

Ertl, J. P. & Schafer, E. W. P. (1967) Cortical activity preceding speech. *Life Sciences* 6:473–79. [EC]

Evans, M. A. & Carr, T. H. (1985) Cognitive abilities, conditions of learning, and the early development of reading skill. *Reading Research Quarterly*. In press. [THC]

Eveleth, P. B & Tanner, J. M. (1976) *Worldwide variation in human growth* Cambridge University Press. [JPR]

Eysenck, H. J. (1939) Review of *Primary mental abilities* by L. L. Thurstone. *British Journal of Educational Psychology* 9:270–75. [taARJ]

(1979) *The structure and measurement of intelligence*. Springer. [HJE]

(1982a) The psychophysiology of intelligence. In: *Advances in personality assessment*, vol. 1, ed. C. D. Spielberger & J. N. Butcher. Erlbaum. [RBC, taARJ]

(1982b) *A model of intelligence*. Springer. [PEV]

(1984) The effect of race on human abilities and mental test scores. In: *Perspectives on bias in mental testing*, ed. C. R. Reynolds & R. T. Brown. Plenum. [HJE]

(1985) The theory of intelligence and the psychophysiology of cognition. In: *Advances in the psychology of human intelligence*, vol. 3, ed. R. J. Sternberg. Erlbaum. [HJE]

Eysenck, H. J. & Barrett, P. (1985) Psychophysiology and the measurement of intelligence. In: *Methodological and statistical advances in the study of*

*individual differences*, ed. C. R. Reynolds & V. Willson. Plenum. [rARJ, EWPS]
Feuerstein, R. (1979) *The dynamic assessment of retarded performers.* University Park Press. [KES]
Fitts, P. M. & Posner, M. I. (1967) *Human performance.* Wadsworth, [YHP]
Floderus-Myrhed, B., Pedersen, N. & Rasmuson, I. (1980) Assessment of heritability for personality, based on a short form of the Eysenck Personality Inventory: A study of 12,898 twin pairs. *Behavior Genetics* 10:153–62. [JPR]
Flynn, J. (1980) *Race, IQ, and Jensen.* Routledge & Kegan Paul. [KES]
(1984) Japanese IQ. *Nature* 308:222. [JPR]
Fraser, I. C. (1984) The psychophysiological measurement of adult intelligence. Thesis, Department of Psychology, University of Edinburgh. [CB]
Frederiksen, J. R. (1980) Component skills in reading: Measurement of individual differences through chronometric analysis. In: *Aptitude, learning, and instruction*, Vol. 1: *Cognitive process analyses of aptitude*, ed. R. E. Snow, P.-A. Federico & W. E. Montague. Erlbaum. [THC]
Freedman, D. G. (1979) *Human sociobiology: A holistic approach.* Free Press. [JPR]
French, J. W., Ekstrom, R. B. & Price, L. A. (1963) *Kit of reference tests for cognitive factors.* Educational Testing Service. [rARJ]
Garrett, H. E., Bryan, A. I. & Perl, R. E. (1935) The age factor in mental organization. *Archives of Psychology*, no. 176. [rARJ]
Gordon, R. A. (1976) Prevalence: The rare datum in delinquency measurement and its implications for the theory of delinquency. In: *The juvenile justice system*, ed. M. W. Klein. Sage. [RAG]
(1984) Digits backward and the Mercer-Kamin law: An empirical response to Mercer's treatment of internal validity of IQ tests. In: *Perspectives on bias in mental testing*, ed. C. R. Reynolds & R. T. Brown. Plenum. [RAG]
Gorsuch, R. L. (1974) *Factor analysis.* W. B. Saunders. [RAG, rARJ]
Gray, J. A. (1982) *The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system.* Oxford University Press. [JPR]
Groen, G. J. & Parkman, J. M. (1972) A chronometric analysis of simple addition. *Psychological Review* 79:329–43. [THC]
Guilford, J. P. (1965) *Fundamental statistics in psychology and education.* 4th ed. McGraw-Hill. [RAG]
Gustafsson, J.-E. (1984) A unifying model for the structure of intellectual abilities. *Intelligence* 8:179–203. [J-EG, rARJ]
Guthrie, J. T. (1973) Models of reading and reading disability. *Journal of Educational Psychology* 65:9–18. [THC]
Haier, R. J., Robinson, D. L., Braden, W. & Williams, D. (1983) Electrical potentials of the cerebral cortex and psychometric intelligence. *Personality and Individual Differences* 4:591–99. [rARJ]
Hansen, J. & Pearson, P. (1983) An instructional study: Improving the inferential comprehension of good and poor fourth-grade readers. *Journal of Educational Psychology* 75:821–29. [KES]
Harlow, H. F. (1959) Learning set and error factor theory. In: *Psychology: A study of a science*, vol. 2, ed. S. Koch. McGraw-Hill. [rARJ]
Harman, H. H. (1960) *Modern factor analysis.* University of Chicago Press. [RAG]
(1967) *Modern factor analysis.* 2nd ed. University of Chicago Press. [taARJ]
Hebb, D. O. (1949) *The organization of behavior.* Wiley. [PEV]
Heber, R. & Garber, H. (1973) The Milwaukee Project: A study of the use of family investigation to prevent cultural–familial mental retardation. In: *Exceptional infant*, vol. 3, ed. B. Z. Friedlander, G. M. Sterritt & G. K. Kirk. Brunner/Mazel. [JSC]
Hemmelgarn, T. E. & Kehle, T. J. (1984) The relationship between reaction time and intelligence in children. *School Psychology International* 5:77–84. [rARJ]
Hendrickson, A. E. (1982) The biological basis of intelligence, Part 1: Theory. In: *A model for intelligence*, ed. H. J. Eysenck. Springer. [CB]
Hendrickson, D. E. (1982) The biological basis of intelligence, Part 2: Measurement. In: *A model for intelligence*, ed. H. J. Eysenck. Springer. [CB]
Hendrickson, D. E. & Hendrickson, A. E. (1980) The biological basis of individual differences in intelligence. *Personality and Individual Differences* 1:3–33. [taARJ, EWPS]
Hennessy, J. J. (1974) *Structure and patterns of mental abilities in several ethnic groups.* Doctoral dissertation, New York University. University Microfilms no. 74–24, 997. [taARJ]
Hennessy, J. J. & Merrifield, P. R. (1976) A comparison of the factor structures of mental abilities in four ethnic groups. *Journal of Educational Psychology* 68:754–59. [RAG, taARJ]
Hogan, D. D. (1971) Cortical response of retardates for AER and audiometry. *American Journal of Mental Deficiency* 75:474. [EC]
Horn, J. L. (1965) Fluid and crystallized intelligence, a factor analytic study of the structure among primary mental abilities. Doctoral dissertation, University of Illinois. [RBC]
(1966) Short period fluctuations in intelligence. University of Denver Research Institute. [RBC]
(1968) Organization of abilities and the development of intelligence. *Psychological Review* 75:242–59. [RBC, J-EG]
(in press) Remodeling old models of intelligence. In: *Handbook of intelligence*, ed. B. Wolman. Prentice-Hall. [JGB]
Hulme, C. (1984) *British Journal of Psychology.* In press. [PMAR]
Humphreys, L. G. (1971) Theory of intelligence. In: *Intelligence: Genetic and environmental influences*, ed. R. Cancro. Grune & Stratton. [rARJ, LVJ]
(1981) The primary mental ability. In: *Intelligence and learning*, ed. M. P. Friedman, J. P. Das & N. O'Connor. Plenum. [taARJ]
(1983) Review of *Ability testing* by A. K. Wigdor & W. R. Garner. *American Scientist* 71:302–3. [taARJ]
(1984) General intelligence. In: *Perspectives on bias in mental testing*, ed. C. R. Reynolds & R. T. Brown. Plenum. [LVJ]
Humphreys, L. G., Fleishman, A. I. & Lin, P.-C. (1977) Causes of racial and socioeconomic differences in cognitive tests. *Journal of Research in Personality* 11:191–208. [taARJ]
Hunt, E. (1976) Varieties of cognitive power. In: *The nature of intelligence*, ed. L. B. Resnick. Erlbaum. [rARJ]
(1978) Mechanics of verbal ability. *Psychological Review* 85:109–30. [PMAR]
(1980) Intelligence as an information-processing concept. *British Journal of Psychology* 71:449–74. [EC]
Hunt, E., Lunneborg, C. & Lewis, J. (1975) What does it mean to be high verbal? *Cognitive Psychology* 7:194–227. [PMAR]
Hunt, J. (1980) *Early development and experience*, vol. 10. 1974 Heinz Werner Lecture Series. [PDB]
Hunt, J. McV. (1961) *Intelligence and experience.* Ronald. [rARJ]
Hunter, J., Jones, L., Vincent, H. & Carmichael, J. W. (1982) Project SOAR: Teaching cognitive skills in a pre-college program. *Journal of Learning Skills* 2:24–26. [AW]
IPAT. (1950, 1959) The Culture Fair Intelligence Test Series. IPAT, Champaign, Ill. [RBC]
Jackson, M. D. & McClelland, J. L. (1979) Processing determinants of reading speed. *Journal of Experimental Psychology: General* 108:151–81. [THC]
Jarman, R. F. (1978) Level I and Level II abilities: Some theoretical re-interpretations. *British Journal of Psychology* 69:257–69. [JPD]
Jensen, A. R. (1966) Social class and perceptual learning. *Mental Hygiene* 50:226–39. [PHS]
(1969) How much can we boost IQ and scholastic achievement? *Harvard Educational Review* 39:1–123. [JPD]
(1971) Do schools cheat minority children? *Educational Research* 14:3–28. [rARJ]
(1973a) *Educability and group differences.* Methuen/Harper & Row. [taARJ, RCN]
(1973b) Level I and Level II abilities in three ethnic groups. *American Educational Research Journal* 4:263–76. [tarARJ]
(1974a) Interaction of Level I and Level II abilities with race and socioeconomic status. *Journal of Educational Psychology* 66:99–111. [tarARJ]
(1974b) How biased are culture-loaded tests? *Genetic Psychology Monographs* 90:185–244. [taARJ]
(1977a) An examination of culture bias in the Wonderlic Personnel Test. *Intelligence* 1:51–64. [taARJ]
(1977b) Cumulative deficit in IQ of blacks in the rural south. *Developmental Psychology* 13:184–91. [PDB, KES]
(1979) *g*: Outmoded theory or unconquered frontier? *Creative Science and Technology* 2:16–29. [taARJ]
(1980a) *Bias in mental testing.* Free Press. [JGB, tarARJ, LVJ, RCN, YHP, PHS]
(1980b) Chronometric analysis of intelligence. *Journal of Social and Biological Structures* 3:103–22. [taARJ, TN]
(1981) Reaction time and intelligence. In: *Intelligence and learning*, ed. M. P. Friedman, J. P. Das & N. O'Connor. Plenum. [taARJ, PMAR]
(1982a) The chronometry of intelligence. In: *Advances in the psychology of human intelligence*, vol. 1, ed. R. J. Sternberg. Erlbaum. [tarARJ, PMAR]
(1982b) Reaction time and psychometric *g*. In: *A model for intelligence*, ed. H. J. Eysenck. Springer. [JB, tarARJ, YHP]
(1983a) The effects of inbreeding on mental ability factors. *Personality and Individual Differences* 4:71–87. [tarARJ, RCJ]

(1983b) The definition of intelligence and factor score indeterminacy. *Behavioral and Brain Sciences* 6:313–15. [taARJ]
(1983c) Again, how much can we boost IQ? Review of *How and how much can intelligence be increased*, ed. D. K. Detterman & R. J. Sternberg. *Contemporary Psychology* 28:756:58. [taARJ]
(1984a) Test validity: *g* versus the specificity doctrine. *Journal of Social and Biological Structures* 7:93–118. [taARJ, AW]
(1984b) The black–white difference on the K-ABC: Implications for future tests. *Journal of Special Education* 18:377–408. [taARJ]
(1984c) Test bias: Concepts and criticisms. In: *Perspectives on bias in mental testing*, ed. C. R. Reynolds & R. T. Brown. Plenum. [rARJ]
(1984d) Sociobiology and differential psychology: The arduous climb from plausibility to proof. In: *Annals of theoretical psychology*, vol. 2, ed. J. R. Royce & L. P. Mos. Plenum. [JPR]
(1985) Race differences and Type II errors: A comment on Borkowski and Krause. *Intelligence* 9:33–39. [rARJ]
(in press) Methodological and statistical techniques for the chronometric study of mental abilities. In: *Methodological and statistical advances in the study of individual differences*, ed. C. R. Reynolds & V. L. Willson. Plenum. [taARJ]
Jensen, A. R. & Figueroa, R. A. (1975) Forward and backward digit span interaction with race and IQ: Predictions from Jensen's theory. *Journal of Educational Psychology* 67:882–93. [tarARJ]
Jensen, A. R. & Inouye, A. R. (1980) Level I and Level II abilities in Asian, white, and black children. *Intelligence* 4:41–49. [rARJ]
Jensen, A. R. & Munro, E. (1979) Reaction time, movement time and intelligence. *Intelligence* 3:121–26. [taARJ]
Jensen, A. R. & Osborne, R. T. (1979) Forward and backward digit span interaction with race and IQ: A longitudinal developmental comparison. *Indian Journal of Psychology* 54:75–87. [taARJ]
Jensen, A. R. & Reynolds, C. R. (1982) Race, social class and ability differences on the WISC-R. *Personality and Individual Differences* 3:423–38. [RAG, tarARJ, KES]
Jensen, A. R., Schafer, E. W. P. & Crinella, F. M. (1981) Reaction time, evoked brain potentials and psychometric *g* in the severely retarded. *Intelligence* 5:179–97. [taARJ]
Jerison, H. J. (1973) *Evolution of brain and intelligence*. Academic Press. [rARJ]
Jones, L. V. (1984) White–black achievement differences: The narrowing gap. *American Psychologist* 39. In press. [LVJ]
Karrer, R. (1984) An analysis of input, central, and motor segments of response time in the mentally retarded. Paper presented at the Conference on Motor Behavior of the Retarded, NICHD, Washington, D.C. [JPD]
Kaufman, A. S. & Kaufman, N. L. (1983) *Kaufman Assessment Battery for Children: Interpretive manual*. American Guidance Service. [RAG, taARJ]
Keele, S. W. (1973) *Attention and human performance*. Goodyear. [rARJ]
Kendall, M. G. & Stuart, A. (1976) *The advanced theory of statistics*, vol. 2, *Inference and relationship*. 3rd ed. Griffin. [taARJ]
Kinchla, R. A. (1974) Detecting target elements in multi-element arrays: A confusability model. *Perception and Psychophysics* 15:149–58. [EC]
Klahr, D. & Wallace, J. G. (1976) *Cognitive development*. Wiley. [THC]
Klapp, S. T., Marshburn, E. A. & Lester, P. T. L. (1983) Short-term memory does not involve the "working memory" of information processing. *Journal of Experimental Psychology: General* 112:240–64. [JPD]
Krech, D., Rosenzweig, M. R. & Bennett, E. L. (1962) Relations between brain chemistry and problem-solving among rats reared in enriched and impoverished environments. *Journal of Comparative and Physiological Psychology* 55:801–7. [PEV]
Lachman, R., Lachman, J. & Butterfield, E. C. (1979) *Cognitive psychology and information processing: An introduction*. Erlbaum. [JPD]
Laosa, L. (1982) School, occupation, culture, and family: The impact of parental schooling on the parent–child relationship. *Journal of Educational Psychology* 74:791–827. [KES]
Laub, J. H. (1983) Urbanism, race, and crime. *Journal of Research in Crime and Delinquency* 2:183–98. [RAG]
Lesser, G. S., Fifer, G. & Clark, D. (1965) Mental abilities of children from different social class and cultural groups. *Monographs of the Society for Research in Child Development* 30, no. 4. [taARJ]
LeVine, R. A. (1975) *Culture, behavior, and personality*. Aldine. [JPR]
Loehlin, J. C., Lindzey, G. & Spuhler, J. N. (1975) *Race differences in intelligence*. W. H. Freeman. [taARJ, JPR]
Luria, A. R. (1976) *Cognitive development: Its cultural and social foundation*. Harvard University Press. [YHP]
Lynn, R. (1978) Ethnic and racial differences in intelligence: International comparisons. In: *Human variation: The biopsychology of age, race, and sex*, ed. R. T. Osborne, C. E. Noble & N. Weyl. Academic Press. [JPR]
(1982) IQ in Japan and the United States shows a growing disparity. *Nature* 297:222–23. [JPR]
MacArthur, R. A. & Elley, W. B. (1963) The reduction of socioeconomic bias in intelligence testing. *British Journal of Educational Psychology* 33:107–19. [RBC]
McCarthy, G. & Donchin, E. (1981) A metric for thought: A comparison of P3 latency and reaction time. *Science* 211:77–80. [EC]
MacGillivray, I., Nylander, P. P. S. & Corney, G. (1975) *Human multiple reproduction*. Saunders. [JPR]
McGue, M., Bouchard, T. J., Jr., Lykken, D. T. & Feuer, D. (1984) Information processing abilities in twins reared apart. *Intelligence* 8:239–58. [rARJ]
McGurk, F. C. J. (1951) Comparison of the performance of Negro and white high school seniors on cultural and noncultural psychological test questions. Catholic University Press [taARJ]
(1953a) On white and Negro test performance and socioeconomic factors. *Journal of Abnormal and Social Psychology* 48:448–50. [taARJ]
(1953b) Socioeconomic status and culturally-weighted test scores of Negro subjects. *Journal of Applied Psychology* 37:276–77. [taARJ]
(1975) Race differences – twenty years later. *Homo* 26:219–39. [taARJ]
McNemar, Q. (1969) *Psychological statistics*. 4th ed. Wiley. [RAG]
Macphail, E. M. (1982) *Brain and intelligence in vertebrates*. Clarendon Press. [EMM]
(1985) Vertebrate intelligence: The null hypothesis. *Philosophical Transactions of the Royal Society of London* B308:37–51. [EMM]
Malina, R. M. (1979) Secular changes in size and maturity: Causes and effects. *Monographs of the Society for Research in Child Development*. Vol. 44, serial no. 179, nos. 3–4. [JPR]
Martin, N. G., Eaves, L. J. & Eysenck, H. J. (1977) Genetical, environmental and personality factors influencing the age of first sexual intercourse in twins. *Journal of Biosocial Science* 9:91–97. [JPR]
Matarazzo, J. D. (1972) *Wechsler's measurement and appraisal of adult intelligence*. 5th ed. Williams & Wilkins. [rARJ, EWPS]
Mercer, J. R. (1984) What is a racially and culturally nondiscriminatory test? In: *Perspectives on bias in mental testing*, ed. C. R. Reynolds & R. T. Brown. Plenum. [RAG, taARJ]
Misawa, G., Motegi, M., Fujita, K. & Hattori, K. (1984) A comparative study of intellectual abilities of Japanese and American children on the Columbia Mental Maturity Scale (CMMS). *Personality and Individual Differences* 5:173–81. [JPR]
Mulaik, S. A. (1972) *The foundations of factor analysis*. McGraw-Hill. [taARJ]
Nagoshi, C. T., Johnson, R. C., DeFries, J. C., Wilson, J. R. & Vandenberg, S. G. (in press) Group differences and first principal component loadings in the Hawaii Family Study of Cognition: A test of the generality of "Spearman's hypothesis." *Personality and Individual Differences*. [RCJ, JRW]
Nettelbeck, T. & Kirby, N. H. (1983) Measures of timed performance and intelligence. *Intelligence* 7:39–52. [taARJ, TN]
Nichols, P. L. (1972) *The effects of heredity and environment on intelligence test performance in 4 and 7 year old white and Negro sibling pairs*. Doctoral dissertation, University of Minnesota. [RAG, taARJ, KES]
Niswander, K. R. & Gordon, M. (1972) *Women and their pregnancies*. W. B. Saunders. [JPR]
Noble, C. E. (1969) Race, reality, and experimental psychology. *Perspectives in Biology and Medicine* 13:10–30. [JGB, rARJ]
(1978) Age, race, and sex in the learning and performance of psychomotor skills. In: *Human variation: The biopsychology of age, race, and sex*, ed. R. J. Osborne, C. E. Noble & N. Weyl. Academic Press. [rARJ]
Occam, William (1320/1965) In: *Gulielmi Ockham opera omnia philosophica et theologica*, ed. E. Moody. Franciscan Institute. [PDB]
Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics) (1982) Profile of American youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery. Department of Defense. [RAG, taARJ]
Ogbu, J. (1982) Societal forces as a context of ghetto children's school failure. In: *The language of children reared in poverty*, ed. L. Feagans & D. Farran. Academic Press. [KES]
Olson, R. K., Kliegel, R., Davidson, B. J. & Foltz, G. (1984) Individual and developmental differences in reading ability. In: *Reading research: Advances in theory and practice*, vol. 4, ed. T. G. Waller & E. MacKinnon. Academic Press. [THC]
Omanson, R. C. (1985) Knowing words and understanding texts. In: *New Directions in Child Development #27: The development of reading skills*, ed. T. H. Carr. Jossey-Bass. In press. [THC]
Ombrédane, A., Robaye, F. & Plumail, H. (1956) Résultats d'une application répétée du matrix-couleur à une population de Noirs Congolai [Results of a repeated administration of the coloured matrices to a population of

Congolese Blacks]. *Bulletin du Centre d'Etudes et Recherches Psychotechniques* 5:129–47. [YHP]
Osborne, R. T. & McGurk, F. C. J., eds. (1982) *The testing of Negro intelligence*, vol. 2. Foundation for Human Understanding. [taARJ]
Pachella, R. G. (1974) The interpretation of reaction time in information-processing research. In: *Human information processing: Tutorials in performance and cognition*, ed. B. H. Kantowitz. Erlbaum. [TN]
Palinscar, A. S. & Brown, A. L. (1984) Reciprocal teaching of comprehension-fostering and monitoring activities. *Cognition and Instruction* 1:117–75. [JGB]
Parmellee, A. & Sigman, M. (1983) Perinatal brain development and behavior. In: *Handbook of child psychology, infancy and developmental psychobiology*, vol. 2, ed. M. M. Haith & J. J. Campos. Wiley. [JSC]
Paul, S. (1985) Speed of information processing: The semantic verification test and general mental ability. Doctoral dissertation, University of California, Berkeley. [rARJ]
Payne, R. B. & Turkat, I. D. (1982) Sex, race, and psychomotor reminiscence. *Bulletin of the Psychonomic Society* 19:336–38. [rARJ]
Pearson, P. (1984) *Handbook of reading research*. Longman. [KES]
Pfefferbaum, A., Wenegrat, B. G., Ford, J. M., Roth, W. T. & Kopell, B. S. (1984) Clinical application of the P3 component of event-related potentials. 2. Dementia, depression and schizophrenia. *Electroencephalography and Clinical Neurophysiology* 59:104–24. [EC]
Platt, J. R. (1964) Strong inference. *Science* 146:347–52. [EC]
Plomin, R. (1983) Developmental behavioral genetics. *Child Development* 54:253–59. [KES]
Polich, J., Howard, L. & Starr, A. (1984) P300 latency correlates with digit span. *Psychophysiology* 20:665–69. [EC]
Poortinga, Y. H. (1971) Cross-cultural comparison of maximum performance tests: Some methodological aspects and some experiments with simple auditory and visual stimuli. *Psychologia Africana Monograph Supplement*, no. 6. [YHP]
(1983) Psychometric approaches to intergroup comparison: The problem of equivalence. In: *Human assessment and cultural factors*, ed. S. H. Irvine & J. W. Berry. Plenum. [YHP]
Posner, M. I. (1966) Components of skilled performance. *Science* 152:1712–18. [taARJ]
(1978) *Chronometric explorations of mind*. Erlbaum. [THC, taARJ]
(1982) Cumulative development of attentional theory. *American Psychologist* 37:168–79. [taARJ]
Posner, M. I. & McLeod, P. (1982) Information processing models: In search of elementary operations. *Annual Review of Psychology* 33:447–514. [THC]
Posner, M. I., Pea, R. & Volpe, B. (1982) Cognitive neuroscience: Developments toward a science of synthesis. In: *Perspectives on mental representations*, ed. J. Mehler, E. Walker & M. Garrett. Erlbaum. [MIP]
Pressey, S. L. & Teter, G. F. (1919) A comparison of colored and white children by means of a group scale of intelligence. *Journal of Applied Psychology* 3:277–82. [taARJ]
Price, G. (1984) Mnemonic support and curriculum selection in teaching by mothers: A conjoint effect. *Child Development* 55:659–68. [KES]
Rabbitt, P. M. A. (1979) Current paradigms and models in human information processing. In: *Human stress and cognition: An information processing approach*, ed. V. Hamilton & D. M. Warburton. Wiley. [TN]
(1981) Cognitive psychology needs models for changes in performance with old age. In: *Attention and Performance* 9, ed. J. Long & A. Baddeley. Erlbaum. [MIP]
& Rodgers, M. (1965) Age and choice between responses in a self-paced repetitive task. *Ergonomics* 8:435–44. [PMAR]
Ramey, C. T., Campbell, F. A., & Finkelstein, N. W. (1984) Course and structure of intellectual development in children at high risk for developmental retardation. In: *Learning and cognition in the mentally retarded*, ed. P. Brooks, R. Sperber, & C. McCauley. Erlbaum.
Reuning, H. (1972) Psychological studies of Kalahari Bushmen. In: *Mental tests and cultural adaptation*, ed. L. J. Cronbach & P. J. D. Drenth. Mouton. [taARJ]
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* 2:175–80. [RAG, taARJ]
Reynolds, C. R. & Jensen, A. R. (1983) WISC-R subscale patterns of abilities of blacks and whites matched on Full Scale IQ. *Journal of Educational Psychology* 75:207–14. [taARJ]
Rock, D. A., Werts, C. E. & Flaugher, R. L. (1978) The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research* 13:403–18. [JGB]
Roth, E. (1964) Die geschwindigkeit der Verarbeitung von Information und ihr Zusammenhang mit Intelligenz. *Zeitschrift für experimentelle und angewandte Psychologie* 11:616–22. [TN]
Rushton, J. P. (1984a) Sociobiology: Toward a theory of individual and group differences in personality and social behavior. In: *Annals of theoretical psychology*, vol. 2, ed. J. R. Royce & L. P. Mos. Plenum. [JPR]
(1984b) Group differences, genetic similarity, and the importance of personality traits. In: *Annals of theoretical psychology*, vol. 2, ed. J. R. Royce & L. P. Mos. Plenum. [JPR]
(1984c) Differential K Theory: The sociobiology of individual differences. *Personality and Individual Differences* 6. [JPR]
Ryan, E. (1981) Identifying and remediating failures in reading comprehension: Toward an instructional approach for poor comprehenders. In: *Reading research: Advances in theory and practice*, vol. 3, ed. T. Waller & G. MacKinnon. Academic Press. [KES]
Salthouse, T. A. & Somberg, B. L. (1982) Skilled performance: Effects of adult age and experience or elementary processes. *Journal of Experimental Psychology: General* 111:176–207. [TN]
Sameroff, A. & Seifer, R. (1983) Familial risk and child competence. *Child Development* 54:1254–68. [KES]
Samuels, S. (1979) The method of repeated readings. *Reading Teacher* 32:403–8. [KES]
Sanders, A. F. (1983) Towards a model of stress and human performance. *Acta Psychologica* 53:61–97. [EC]
Sandoval, J. (1982) The WISC-R factoral validity for minority groups and Spearman's hypothesis. *Journal of School Psychology* 20:198–204. [RAG, taARJ]
Sandoval, J. & Millie, M. P. W. (1980) Accuracy of judgments of WISC-R item difficulty for minority groups. *Journal of Consulting and Clinical Psychology* 48:249–53. [taARJ]
Scarr, S. (1981a) Implicit messages: A review of *Bias in mental testing*. *American Journal of Education* 89:330–38. [KES]
(1981b) *Race, social class, and individual differences in IQ*. Erlbaum. [RAG, taARJ, KES]
Scarr, S., Caparulo, B. K., Ferdman, B. M., Tower, R. B. & Caplan, J. (1983) Developmental status and school achievements of minority and non-minority children from birth to 18 years in a British Midlands town. *British Journal of Developmental Psychology* 1:31–48. [JPR]
Scarr, S. & Carter-Saltzman, L. (1982) Genetics and intelligence. In: *A handbook of human intelligence*, ed. R. J. Sternberg. Cambridge University Press. [CB]
Schafer, E. W. P. (1982) Neural adaptability: A biological determinant of behavioral intelligence. *International Journal of Neuroscience* 17:183–91. [taARJ, EWPS]
(1984) Habituation of evoked cortical potentials correlates with intelligence. *Psychophysiology* 21:597. [EWPS]
Schafer, E. W. P. & Marcus, M. M. (1973) Self-stimulation alters human sensory brain responses. *Science* 181:175–77. [EWPS]
Schechter, G. & Callaway, E. (1984) Attention selectively modulates parallel visual processes. Abstract. Society for the Neurosciences, 14th Annual Meeting, vol. 10, part 2. [EC]
Schiff, M., Duyme, M., Dumaret, A. & Tomkiewicz, S. (1982) How much *could* we boost scholastic achievement and IQ scores? A direct answer from a French adoption study. *Cognition* 12:165–96. [KES]
Schmid, J. & Leiman, J. M. (1957) The development of hierarchical factor solutions. *Psychometrika* 22:53–61. [taARJ]
Schönemann, P. H. (1981) Factorial definitions of intelligence: Dubious legacy of dogma in data analysis. In: *Multidimensional data representations: When and why*, ed. I. Borg. Mathesis Press. [PHS]
(1983) Do IQ tests really measure intelligence? *Behavioral and Brain Sciences* 6:311–15. [PHS]
Schwartz, R. M. & Stanovich, K. E. (1981) Flexibility in the use of graphic and contextual information by good and poor readers. *Journal of Reading Behavior* 13:263–69. [THC]
Sen, A., Jensen, A. R., Sen, A. K. & Arora, I. (1983) Correlation between reaction time and intelligence in psychometrically similar groups in America and India. *Applied Research in Mental Retardation* 4:139–52. [taARJ]
Serpell, R. (1979) How specific are perceptual skills? *British Journal of Psychology* 70:365–80. [YHP]
Shucard, D. & Horn, J. (1972) Evoked cortical potentials and measurement of human abilities. *Journal of Comparative and Physiological Psychology* 78:59–68. [taARJ]
Shuey, A. M. (1966) *The testing of Negro intelligence*. 2nd ed. Social Science Press. [taARJ]
Silverstein, A. B. (1980a) Estimating the general factor in the WISC and the WAIS. *Psychological Reports* 46:189–90. [taARJ]

(1980b) Estimating the general factor in the WISC-R. *Psychological Reports* 47:1185–86. [taARJ]
Sisco, F. H. (1982) Sex differences in the performance of deaf children on the WISC-R Performance Scale. Doctoral dissertation, University of Florida. University Microfilms no. DA 8226432. [taARJ]
Sharp, D. M. (1984) Inspection time, decision time and visual masking. Thesis, Department of Psychology, University of Aberdeen. [CB]
Sherry, D. F. (1984) What food-storing birds remember. *Canadian Journal of Psychology* 38:304–21. [EMM]
Sörbom, D. (1974) A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology* 27:229–39. [J-EG]
Spearman, C. (1904) General intelligence, objectively determined and measured. *American Journal of Psychology* 15:201–93. [RBC]
(1923) *The nature of "intelligence" and the principles of cognition.* Macmillan. [tarARJ]
(1927) *The abilities of man.* Macmillan. [tarARJ, RCN]
Stankov, L. (1983) The role of competition in human abilities revealed through auditory tests. *Multivariate Behavioral Research.* Monograph no. 83-1. [taARJ]
Stanovich, K. E. & West, R. F. (1979) The effect of orthographic structure on the word search performance of good and poor readers. *Journal of Experimental Child Psychology* 28:258–67. [THC]
Stanovich, K. E., West, R. F. & Feeman, D. (1981) A longitudinal study of sentence context effects in second-grade children. *Journal of Experimental Child Psychology* 32:185–99. [THC]
Sternberg, R. J. (1977) *Intelligence, information processing and analogical reasoning: The componential analysis of human abilities.* Erlbaum. [PK]
(1982) *Handbook of human intelligence.* Cambridge University Press. [HJE, PMAR]
(1984a) *Beyond IQ: A triarchic theory of intelligence.* Cambridge University Press. [JGB]
(1984b) Toward a triarchic theory of human intelligence. *Behavioral and Brain Sciences* 7:269–87. [THC, KES]
Sternberg, R. J., Conway, B. E., Ketron, J. R. & Bernstein, M. (1981) People's conceptions of intelligence. *Journal of Personality and Social Psychology: Attitudes and Social Cognition* 41:37–55. [taARJ]
Sternberg, R. J. & Gardner, M. K. (1982) A componential interpretation of the general factor in human intelligence. In: *A model for intelligence*, ed. H. J. Eysenck. Springer. [tarARJ]
Sternberg, R. J. & Powell, J. S. (1983) Comprehending verbal comprehension. *American Psychologist* 38:878–93. [taARJ]
Sternberg, S. (1966) High speed scanning in human memory. *Science* 153:652–54. [PMAR]
(1969) The discovery of processing stages: Extensions of Donders' method. In: *Attention and performance II*, W. G. Koster. *Acta Psychologica* 30:276–315. [THC, PMAR]
Straumanis, J. J., Jr., Shagass, C. & Overton, D. A. (1973) Auditory evoked response in young adults with Down's syndrome and idiopathic mental retardation. *Biological Psychiatry* 6:75–90. [EC]
Teichner, W. H. & Krebs, M. J. (1974) Laws of visual choice reaction time. *Psychological Review* 81:75–98. [TN]
Thomas, B. (1984) Early toy preferences of four-year-old readers and nonreaders. *Child Development* 55:424–30. [KES]
Thurstone, L. L. (1938) *Primary mental abilities.* University of Chicago Press. [RCN]
Tierney, R. & Cunningham, J. (1984) Research on teaching reading comprehension. In: *Handbook of reading research*, ed. P. Pearson. Longman. [KES]
Trotman, F. (1977) Race, IQ, and the middle class. *Journal of Educational Psychology* 69:266–73. [KES]
Tulkin, S. (1968) Race, class, family, and school achievement. *Journal of Personality and Social Psychology* 9:31–37. [KES]
Undheim, J. O. (1981a) On intelligence. 1. Broad ability factors in 15-year-old children and Cattell's theory of fluid and crystallized intelligence. *Scandinavian Journal of Psychology* 22:171–79. [rARJ]
(1981b) On intelligence. 2. A neo-Spearman model to replace Cattell's theory of fluid and crystallized intelligence. *Scandinavian Journal of Psychology* 22:181–87. [J-EG, rARJ]
(1981c) On intelligence. 4. Toward a restoration of general intelligence. *Scandinavian Journal of Psychology* 22:251–65. [rARJ]
U.S. Department of Labor, Manpower Administration (1970) *Manual for the USES General Aptitude Test Battery.* U.S. Employment Service. [RAG, taARJ]
Van de Vijver, F. J. R. & Poortinga, Y. H. (1982) Cross-cultural generalization and universality. *Journal of Cross-Cultural Psychology* 13:387–408. [YHP]
Vavrus, L. G., Brown, T. L. & Carr, T. H. (1983) Component skill profiles of reading ability: Variations, tradeoffs, and compensations. Paper presented at the meeting of the Psychonomic Society, November 1983, San Diego, Calif. [THC]
Vernon, P. A. (1981a) Level I and Level II: A review. *Educational Psychologist* 16:45–64. [taARJ]
(1981b) Reaction time and intelligence in the mentally retarded. *Intelligence* 5:345–55. [taARJ]
(1983) Speed of information processing and general intelligence. *Intelligence* 7:53–70. [tarARJ, TN, PMAR]
Vernon, P. A. & Jensen, A. R. (1984) Individual and group differences in intelligence and speed of information processing. *Personality and Individual Differences* 5:411–23. [tarARJ, TN]
Vernon, P. E. (1979) *Intelligence, heredity, and environment.* Freeman. [PMAR]
(1982) *The abilities and achievements of Orientals in North America.* Academic Press. [JPR]
Veroff, J., McClelland, L. & Marquis, K. (1971) Measuring intelligence and achievement motivation in surveys. Final report to U.S. Dept. of HEW, OEO, contract no. OEO-4180. Survey Research Center, Institute for Social Research, University of Michigan. [taARJ]
Viaud, G. (1960) *Intelligence: Its evolution and forms.* Hutchinson. [rARJ]
Wade, M., Hoover, J. & Newell, K. (1984) Training reaction and movement times of moderately and severely mentally retarded persons in aiming movements. *American Journal of Mental Deficiency* 89:174–79. [KES]
Wallach, M. & Wallach, L. (1979) Helping disadvantaged children learn to read by teaching them phoneme identification skills. In: *Theory and practice of early reading*, vol. 3, ed. L. Resnick & P. Weaver. [KES]
Wechsler, D. (1958) *The measurement and appraisal of adult intelligence.* 4th ed. Williams & Wilkins. [taARJ]
Weinrich, J. D. (1977) Human sociobiology: Pair bonding and resource predictability (Effects of social class and race). *Behavioral Ecology and Sociobiology* 2:91–118. [JPR]
Weiss, V. (1984) Psychometric intelligence correlates with interindividual different rates of lipid peroxidation. *Biomedica Biochimica Acta* 6:755–63. [EC]
Welford, A. T. (1980) *Reaction times.* Academic Press. [YHP]
Werner, H. & Kaplan, E. (1952) The acquisition of word meanings: A developmental study. *Monographs of the Society for Research in Child Development*, no. 51. [taARJ]
Wherry, R. J. (1959) Hierarchical factor solutions without rotation. *Psychometrika* 24:45–51. [taARJ]
Whimbey, A. (1975) *Intelligence can be taught.* Dutton. [AW]
(1981) *Competency in comprehension (and scientific literacy) as reasoning.* Washington, D.C.: Fund for the Improvement of Post-secondary Education. [AW]
(1983) *Analytical reading and reading.* Innovative Sciences. [AW]
(1984) *Mastering reading through reasoning.* Innovative Sciences. [AW]
Whimbey, A. & Lochhead, J. (1982) *Problem solving and comprehension: A short course in analytical reasoning.* Franklin Institute Press. [AW]
Willerman, L. (1973) Activity level and hyperactivity in twins. *Child Development* 44:288–93. [JPR]
Williams, J. (1980) Teaching decoding with an emphasis on phoneme analysis and phoneme blending. *Journal of Educational Psychology* 72:1–15. [KES]
Wilson, E. O. (1975) *Sociobiology: The new synthesis.* Harvard University Press. [JPR]
Wilson, R. S. (1983) The Louisville Twin Study: Developmental synchronies in behavior. *Child Development* 54:298–316. [JPR]
Wilson, R. & Matheny, A. (1983) Mental development: Family environment and genetic influences. *Intelligence* 7:195–215. [KES]
Wober, M. (1974) Towards an understanding of the Kiganda concept of intelligence. In: *Culture and cognition: Readings in cross-cultural psychology*, ed. J. W. Berry & P. R. Dasen. Methuen. [JSC]
Yerkes, R. M. & Dodson, J. D. (1908) The relation of strength of stimulus to rapidity of habit formation. *Journal of Comparative Neurology and Psychology* 18:458–82. [rARJ]