Journal of Vocational Behavior **29**, 301–331 (1986)

MAJOR CONTRIBUTIONS

g: Artifact or Reality?

ARTHUR R. JENSEN

School of Education, University of California, Berkeley

The highest common factor in any large and diverse collection of mental tests is measured by means of factor analysis, and is conventionally labeled psychometric *g* (for general ability). The *g* factor, which is highly correlated across even quite different batteries of tests, provided the tests are fairly numerous and varied, reflects the empirical fact of *positive manifold*, that is, positive correlations between all mental tests. After briefly explicating the general psychometric conditions and factor analytic methods for the measurement of *g*, this article addresses the theoretically important question of whether *g* is merely an artifact of the method of constructing psychometric tests and the mathematical operations of factor analysis or whether it has an authentic claim to represent some natural phenomenon that exists independently of psychometrics and factor analysis. Several lines of evidence which refute the argument that *g* is a methodological artifact are presented. The *g* factor, far more than any other linearly independent sources of variance in psychometric tests, is correlated with various phenomena that are wholly independent of both psychometrics and factor analysis, such as the heritability of test scores, familial correlations, the effects of inbreeding depression and of hybrid vigor, evoked electrical potentials of the brain, and reaction times to elementary cognitive tasks which have virtually no intellectual content. This evidence of biological correlates of *g* supports the theory that *g* is not a methodological artifact but is, indeed, a fact of nature. However, the causal nature of *g* itself is not yet scientifically established. That goal must await further advances in neuroscience. © 1986 Academic Press, Inc.

The hypothesis of general mental ability, in which human individual differences range widely, was first formally propounded by Sir Francis Galton (1869). Galton's hypothesis was not subjected to rigorous empirical scrutiny, however, until there was developed a methodology adequate to the task. The great pioneer in this development was Charles Spearman (1904, 1927), whose invention of factor analysis made the construct of

Presented at the fall conference of the Personnel Testing Council of Southern California, "The *g* Factor in Employment Testing," Newport Beach, CA, October 10, 1985. Requests for reprints should be sent to Arthur R. Jensen, School of Education, University of California, Berkeley, CA 94720.

0001-8791/86 $3.00

general ability the subject of some 80 years of empirical inquiry and controversy in the field of psychometrics.

In the realm of mental tests, general ability is referred to more specifically as psychometric *g*, or more briefly as just *g*, the designation originated by Spearman. Spearman's *g*, however, because of its intimate connection with mental tests and the mathematical operations of factor analysis, became a rather narrower conception of general ability than Galton's notion. Galton conceived of general ability more broadly in essentially biological and evolutionary terms. But Galton's view faded into the background as the theory of general ability became exclusively identified with the *g* derived from the factor analysis of mental tests by Spearman and his many successors in psychometric research on individual differences.

The exclusive dependence on conventional psychometric tests and on the complex mathematical technology of factor analysis as the basis of the argument for the existence of *g* has given rise to one of the most fundamental and contentious questions in this field. It is this: Is *g* merely a methodological artifact, that is, merely a product of psychometric testing and the mathematical manipulations of applying factor analysis to the intercorrelations of various tests? Or does the *g* revealed by factor analysis reflect a reality that exists independently of psychometric tests and factor analysis? Virtually no one today disputes that a *g* factor can be extracted from the correlations among any large and diverse collection of mental ability tests, and that the *g* factor is usually substantial in the sense of subsuming a relatively large proportion of the total variance in all of the tests as compared with other factors besides *g*. The point that is being questioned is whether the *g* factor represents any reality outside the operations of psychometric tests and factor analysis. Is *g* actually the Galtonian notion of general ability as a biological reality, or is this concept properly restricted to its more limited Spearmanian or exclusively psychometric meaning? Before we can even begin to examine this question, we should review some of the well-established facts about *g* strictly within its own realm of psychometrics and the factor analysis of conventional mental tests.

An item is the elemental unit of a mental test. An *item* is a specific mental task to which a person's overt response can be objectively scored, that is, classified or quantified (e.g., "right" or "wrong" = 1 or 0), graded on a scale (e.g., "poor," "fair," "good," "excellent" = 0, 1, 2, 3), counted (e.g., number of digits recalled, number of parts of a puzzle fitted together within a given time limit), or measured on a ratio scale (e.g., the time interval between presentation and completion of a task). The scoring is *objective* in the same sense that all scientific measurement is objective; that is, there is a high degree of agreement among all competent observers making the measurements. Objective measurement may depend on special instruments or special training of the observers. A task is said to be a *mental* task if variance (i.e., individual differences) in performance is negligibly attributable to individual differences in sheer physical capacities, such as sensory acuity or muscular strength, in the population of interest. A task qualifies as an appropriate mental test item only if the testee understands the requirements of the task, through related prior experience, preliminary instructions by the tester, or practice on easy examples with informative feedback as to the correctness of the testee's performance. For an item to be psychometrically useful in a test, its variance must be greater than zero in the population of interest; that is, there must be nonchance individual differences in scores on the item. Items that compose tests of *ability* (as contrasted with personality, attitude, and interest inventories) are also characterized by the property that the items are objectively scorable in terms of the *goodness* of the testee's performance, simply in the sense that there is universal agreement that, say, the answer "4" to the question, "What is 2 plus 2?" is *better* than the answer "5" (or some other number besides "4"), or that solving a puzzle in 2 min is *faster* than solving it in 3 min, or that recalling 7 digits indicates a *larger* memory span than recalling only 5 digits. These judgments per se do not concern the social, practical, or moral value of the particular performance. All test items are conventionally scored so that "goodness" of performance is always represented by a higher score.

A test is composed of a number of items. A *test* may be composed of any finite number of items of any degree of diversity involving different sensory and response modalities, different media (words, numbers, symbols, pictures of familiar things, objects), different types of task requirements (discrimination, generalization, recall, naming, verbal expression, manipulation of objects, comparison, decision, inference, etc.), and variation in task complexity ranging all the way from simple reaction time to inductive and deductive reasoning. The number and variety of items in a test are governed by the test constructor's purpose and the practical limitations and cost/benefit ratio for the use of the test in a given setting.

Single items show generally low but positive correlations with one another when administered to large representative samples of the general population. Single test items measure very little in common with other single items. Most of the variance on a single item is unique to itself, that is, it is not correlated with whatever is measured by other test items. This is clearly evident from the fact that the interitem correlations in standard tests are seldom as high as .20 and are usually closer to .10. Even in a test with a high degree of item homogeneity (i.e., similarity of item types), the interitem correlations are surprisingly low. In a large random sample of school children, for example, the Ravens Standard Progressive Matrices, which probably has greater item homogeneity than any other standard intelligence test, shows an average item intercorrelation of only +.13.

The saving grace is the fact that mental test items of virtually all kinds are positively correlated with one another in the general population.

Negative and zero correlations are almost entirely due to sampling error. As sample size increases, the negative and zero correlations decrease to the vanishing point. The fact of ubiquitous positive correlations between items means they are all measuring something in common, and the larger the number of items, the more of this common factor is measured by the aggregate. If, in the collection of items that compose a test, the single-item scores are summed for each person, we obtain the individual's *raw score* on the test. The total variance of raw scores on the test in the population is equal to the sum of all the single-item variances plus twice the sum of all the item covariances. Since, for n item variances, there are $n(n - 1)$ item covariances, increasing the number of items in a test increases the total item covariance at a greater rate than it increases the total item variance. The covariance divided by the total variance is the internal consistency reliability of the test, or the proportion of the total variance attributable to whatever it is that all of the items measure in common. For most standardized tests, this value is generally above .90. But *any* collection of ability items, however diverse, will yield a similar value, or any value one would like, less than 1, provided a sufficient number of items is included in the collection.

In any large collection of diverse items, the items can be clustered in terms of their intercorrelations, grouping various items with the highest intercorrelations together to form smaller, more homogeneous, sets of items called *subtests*. Such subtests are usually composed of quite similar item types, such as vocabulary items, numerical items, figural items, and so forth. Such relatively homogeneous tests can be made to have as high internal consistency reliability as one would like simply by including more items of the same type. Thus the internal consistency reliability of a test is a function of two effects: the average item intercorrelation and the number of items.

All varieties of mental ability tests are positively correlated with one another in the general population. Diverse tests, assuming they are composed of enough items to ensure high internal consistency reliability, always show nonzero positive correlations when administered to large, unbiased samples of the population. The sizes of the correlations may range widely, from near zero to over .90, depending on the diversity of the tests, and the average correlation may differ accordingly. But the really important fact, which by now has the status of a fact of nature, is that the correlations are all positive—a phenomenon termed *positive manifold*—regardless of the diversity of the tests, provided they are mental *ability* tests, as previously defined, and also have adequate reliability (since any two tests cannot be more highly correlated than the [geometric] mean of their respective reliability coefficients). Apparent violations of positive manifold may be observed when a battery of tests is administered to samples that are markedly biased with respect to abilities. In the general population, for example, verbal tests and numerical tests are very highly correlated. But when such tests are given to a group composed of equal numbers of highly selected university students in law and in engineering, the correlation between verbal and numerical tests may be close to zero or may even be a negative correlation, because law students, on average, tend to be relatively high on verbal and low on numerical, while engineering students show the opposite pattern.

No one has yet been able to devise a number of different ability tests which, when correlated with one another in a large and representative sample of the general population, do not show positive manifold. The leading American psychometrician, L. L. Thurstone, spent many years trying to devise tests that he hoped would afford pure measures of a number of supposedly distinct abilities, such as verbal, numerical, spatial, reasoning, and memory. No matter how refined and homogeneous these various tests were made, they always displayed substantial positive correlations with one another, indicating that all of these tests measured something in common—a *general factor*—in addition to whatever special ability was uniquely measured by each test—abilities that Thurstone termed the *primary mental abilities*. Thurstone's tests of "primary mental abilities" each measured a single general factor common to all of the tests in addition to the particular primary ability each test was specifically designed to measure. It is now amply apparent that it would be impossible to have it otherwise. The phenomenon of positive manifold is about as inexorable as gravitation.

The correlation of each of a number of tests in a battery of tests with the general factor common to all of the tests can be determined by the technique of factor analysis. Factor analysis is essentially a class of mathematical techniques for converting a number of observed variables (e.g., test scores) into a usually much smaller number of hypothetical variables, called *factors,* which together represent all or most of the variance that any of the observed variables have in common, referred to as *common factor variance*. The total variance in all of the observed variables is composed of the common factor variance and the sum of all the variances that are *unique* to each of the variables. The common factor variance may be composed of one or more factors, depending on the nature of the variables. The factors may be uncorrelated with one another (*orthogonal* factors) or correlated with one another (*oblique* factors), depending on the method of factor analysis. Thus, by means of factor analysis one can partition the total variance on an observed variable into various hypothetical components consisting of one or more factors and the variable's *uniqueness,* which is the variance of a single variable that it does not have in common with any other variable in the set of factor-analyzed variables. The *uniqueness* of a given observed variable consists of two parts: (1) the reliable or true-score variance that is unique to the observed variable, which is termed the *specificity* of the given variable, and (2) the unreliability or *error* variance in the given variable.

The correlation between an observed variable and a particular hypothetical factor is termed the *factor loading* (or, less commonly, *factor saturation*) of the variable on the particular factor. The squared factor loading is the proportion of the total variance in the observed variable that is "accounted for" by the factor. The sum of an observed variable's squared factor loadings is termed the variable's *communality*, or the variable's total common factor variance, conventionally symbolized as h^2. (The symbol h^2 for *communality* should never be confused with *heritability*, which is also symbolized as h^2. Heritability refers to the proportion of the total variance in phenotypes that is attributable to genetic factors. There is no theoretical connection between communality and heritability, and the fact that both concepts share the same symbol, h^2, is merely an unfortunate coincidence.)

Just as the matrix of correlations between observed variables can be factor analyzed, so too can the correlations between three or more oblique (i.e., correlated) factors, thereby yielding one or more *higher order factors*. Hence factors can be represented as a hierarchy in terms of their degree of generality, going from first-order (or *primary*) factors, to second-order factors, and so on. The highest order factor at the apex of the hierarchical factor structure is the *general factor*, which, following Spearman, is conventionally labeled *g* (always a lowercase *g*) when the observed variables entering into the factor analysis are scores on a wide variety of tests of mental abilities. A hierarchical factor structure is illustrated in Fig. 1. The connecting lines represent correlations. Each higher level in this hierarchical structure is more general than the lower level. Variance that is unique to each of the tests is "filtered out" at the level of the *primary factors;* variance that is unique to each of the primary factors is "filtered out" at the level of second-order factors, and so on. The *g* factor is the

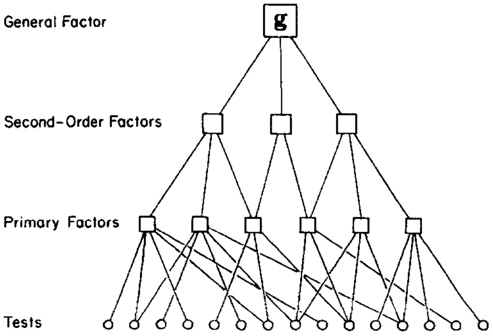

FIG. 1. Example of a hierarchical factor analysis with three levels.

highest degree of generality. Factors below the general factor in the hierarchy are also referred to as *group factors,* because their variance is shared by only certain groups of tests. Prominent group factors are *verbal, spatial,* and *numerical.*

The number of levels in the hierarchy and the number of factors at each level are mainly a function of the number and diversity of the tests that are factor analyzed. When there are relatively few tests, *g* emerges as a second-order factor. A hierarchical factor analysis of the 12 subtests of the Wechsler Intelligence Scale for Children (WISC), for example, yields three primary factors (verbal, spatial, memory) and only one second-order factor (*g*) (Jensen & Reynolds, 1982). Combining the 13 subtests of the Kaufman Assessment Battery for Children (K-ABC) with the 12 WISC subtests yields the very same factor structure (Naglieri & Jensen, in press).

In an *orthogonalized* hierarchical factor analysis (Schmid & Leiman, 1957; Wherry, 1959), each of the factors is uncorrelated with every other factor, both within and between all levels of the hierarchy. But the final outcome of the analysis yields the loadings (i.e., correlations) of each of the tests on each of the uncorrelated factors at each level of the hierarchy. In a factor analysis of ability tests, the *g* factor typically accounts for more of the total variance than any of the group factors and often accounts for a larger proportion of the total variance in the tests than is accounted for by all of the group factors combined.

Although there are a number of methods of nonhierarchical factor analysis in which only the primary factors are extracted, there is now a high degree of consensus among researchers studying abilities that a hierarchical factor analysis provides the best representation of the correlational structure of human abilities. The first *principal component* of a correlation matrix can also represent the *g* factor and is usually very highly correlated with the hierarchical *g*. But tests' loadings on the first principal component are slightly contaminated by some admixture of each test's unique variance in its loading on a principal component. The first *principal factor* of a correlation matrix excludes the unique variance and is therefore preferable to the first principal component as a measure of *g*. But both methods are alike in having two main disadvantages: (1) They are more strongly affected than is a hierarchical *g* by *psychometric sampling,* that is, the particular combination and number of the various types of tests included in the analysis; and (2) under freakish circumstances, which are rare in the abilities domain, they can spuriously create the appearance of a general factor in a collection of variables in which there is in fact no real general factor and in which a hierarchical analysis would yield no general factor at all. As an extreme but clear-cut example, consider, say, 10 variables, among which the set of Variables 1–5 are highly intercorrelated and the set of Variables 6–10 are highly intercorrelated, but all the variables in the first set have zero correlations with all of the variables in the second set. The first principal component and the first principal factor will both show fairly large positive loadings on all 10 variables, when there is obviously no general factor that is common to all 10 variables. If there existed a true general factor, there should be no correlations of zero between any of the variables. A hierarchical factor analysis applied to the same sets of correlations described above could not yield a general factor; it could yield only a number of primary factors or primary factors and two or more higher order factors. But the hierarchy would be truncated, without a *g* factor at the apex. In actual fact, however, I have yet to find a collection of psychometric tests for which the first principal component, the first principal factor, and the hierarchical *g* are not almost perfectly correlated, with intercorrelations typically above .95 and usually close to .99.

The g factor is quite stable across different collections of diverse mental ability tests. Any limited collection of tests may be regarded as a sample of the universe of all tests. Therefore, the statistical characteristics of any limited collection of tests will not perfectly represent the corresponding parameters of the universe of tests. In brief, there will be psychometric sampling error. The *g* factor, by any method of extraction, is subject to this source of error. The *g* extracted from one battery of tests will not be exactly the same *g* extracted from a different battery of tests. A necessary corollary is that a given test will not show exactly the same *g* loading when factor analyzed in different batteries of tests. In brief, the *g* factor and the *g* loading of any particular test are not invariant across different samples of tests. This fact per se does not undermine the construct of *g*. Some degree of error is ubiquitous in *all* measurement, and this is true in *every* empirical science. Inevitable error simply calls for proper assessment. If the *g* of any battery of tests bore no resemblance to the *g* of any other battery, then, of course, *g* would have little, if any, scientific interest and would hardly qualify as an important theoretical construct in the theory of human ability. But, in fact, quite the opposite is the case.

The *g* factor is remarkably stable across different collections of mental tests, even collections of tests that bear hardly any superficial resemblance to one another. For example, the *g* of just the six verbal tests of the Wechsler Adult Intelligence Scale (WAIS) and the *g* of just the six performance tests are correlated .80. The *g* of a battery of six diverse tests of immediate or short-term memory (paired associates, meaningful prose, free recall of words, digit span, memory for forms, memory for objects) was found to have a correlation of +.87 with the *g* of four quite different tests (motor speed, vocabulary, arithmetic, form board) (Garrett, Bryan, & Perl, 1935). In the most recent and probably most rigorous and large-scale study of the stability of *g* across different test batteries,

R. L. Thorndike (in press) made use of 65 highly diverse tests used in the armed services and administered to a large sample of enlisted personnel. First, Thorndike made up at random 6 nonoverlapping batteries of 8 tests each. Then, 17 diverse "target" tests were each singly included in each of the 6 test batteries and the *g* factor (as represented by the first principal factor) was extracted from the total of 9 tests in each battery. Hence there were obtained 6 *g* loadings for each of the 17 target tests, resulting from including each of the target tests in each of the 6 nonoverlapping batteries of diverse tests. The average correlation between the 17 *g* loadings across any two batteries was +.83. In other words, a given test was relatively invariant in its *g* loading despite considerable variation between the 6 different test batteries in which it was factor analyzed. A test's composite *g* loading, that is, the average of its *g* loadings in all 6 batteries, would reflect less psychometric sampling error than any single *g* loading. Thus the composite *g* loading should asymptotically approach the test's "true" *g* loading as we increase the number of different test batteries. Just as we can speak of a hypothetical "true score" on a test, we can speak of a hypothetical "true *g*." And just as the obtained score on a test asymptotically approaches its hypothetical true score as a function of the number of items in the test, so, too, the obtained *g* factor of a battery of tests asymptotically approaches the hypothetical true *g* as we increase the number of tests entering into the factor analysis. In Thorndike's study, with just 6 test batteries, each consisting of 8 tests besides the target tests, it can be shown that the correlation of the mean of the 6 *g* loadings of each of the 17 tests is correlated +.98 with the tests' hypothetical true *g* loadings. If larger test batteries had been used, the consistency of *g* across batteries would be even higher. Also, Thorndike used as the estimate of *g* the first principal factor, which is always more sensitive to psychometric sampling variation and therefore is less stable than is a hierarchical *g*.

It is evident from these findings that in the context of psychometric tests and factor analysis, the *g* factor is a highly ubiquitous phenomenon and its measurement is highly stable, even across diverse batteries of tests. Spearman (1927) summarized this fact in his famous "theorem" of "the indifference of the indicator" of *g* (p. 197).

At present g is known only by its site, not by its nature. Spearman (1927) stated:

> This general factor *g*, like all measurements anywhere, is primarily not any concrete thing but only a value or magnitude. Further, that which this magnitude measures has not been defined by declaring what it is like, but only by pointing out where it can be found. It consists in just that constituent—whatever it may be—which is common to all the abilities inter-connected by the tetrad equation [i.e., Spearman's method for identifying the *g* factor in a battery of tests]. This way of indicating what *g* means is just as definite as when one indicates a card by staking on the

> back of it without looking at its face. Such a defining of *g* by site rather than by nature is what was meant originally when its determination was said to be only "objective." Eventually, we may or may not find reason to conclude that *g* measures something that can appropriately be called "intelligence." Such a conclusion, however, would still never be a definition of *g*, but only a "statement about it." (Spearman, 1927, pp. 75–76)

Spearman's statement is still valid today, if we remain only within the confines of psychometrics. We can note differences in the *g* loadings of various tests and try to discern the features that distinguish between high- and low-*g*-loaded tests. When Spearman made such comparisons of more than 100 various tests he had factor analyzed, he concluded that *g* is most strongly represented in tests that involve the "eduction of relations and correlates" and "abstraction." A test's relative standing on *g* could not be inferred from its superficial characteristics, such as the sensory or response modality involved, whether verbal or nonverbal, numerical or figural, paper-and-pencil test or performance test, or other formal features. Vocabulary and block design, for example, are highly dissimilar tests in appearance and task requirements, yet they are the 2 most highly *g*-loaded tests of all the 12 tests in the Wechsler battery. Among various psychometric test items, in general, the size of the *g* factor seems to reflect the amount or complexity of the mental manipulation, or cognitive processing, required for the testee to arrive at the correct response. A clear example of this is the fact that forward digit span has only about half as large a *g* loading as backward digit span, when both subtests are factor analyzed among the 11 other subtests of the WISC (Jensen & Figueroa, 1975).

g is the sine qua non of all intelligence tests. All so-called intelligence tests, or "IQ" tests, even when they have not been constructed with reference to factor analysis, are found to be very highly *g* loaded. Yet the average correlation between total scores on various standardized IQ tests in representative samples of the general population is less than perfect—about +.80, or +.90 when corrected for attenuation (Jensen, 1980, pp. 315–316). The main reason for the lack of perfect correlation, besides unreliability, is that various IQ tests, although all are highly *g* loaded, also reflect differing amounts of variance attributable to various non-*g* group factors, such as verbal, spatial, and memory factors, as well as reliable nonfactor variance that is specific to each test.

For scientific purposes it is probably best to identify the concept of intelligence with *g*. Otherwise, as Spearman pointed out, there is no possibility for an objective criterion for determining whether a given test or battery of tests provides a better or poorer measure of intelligence than some other test. To identify intelligence as the totality of *all* mental abilities is a conceptual muddle. Intelligence is not the whole of mental ability; besides *g* there is some indefinite number of primary or group factors independent of *g*. Hence the construct of intelligence can be most precisely distinguished from other abilities by means of factor analysis and should not be a label for just *any* kind of ability in which we can observe individual differences. It seems sensible to identify the term *intelligence* with *g*, because *g* is the highest common factor in any large and diverse collection of tests of various abilities. But there is also another good reason to identify intelligence with *g*. The *g* factor is more highly correlated than any other factors (independent of *g*) with individual differences in those observable behaviors that are most commonly associated with the use of the word *intelligence* in popular parlance.

The practical predictive validity of psychometric tests is mainly dependent on their g loading. Many different tests have substantial and practically useful predictive validity for performance in school, college, in the armed services training programs, and in hundreds of different occupations in business, industry, and the civil service. My examination of the correlational evidence for the validity of tests in these settings leads me to the conclusion that virtually all test validity would be drastically reduced, usually to a level of practical uselessness, if the *g* factor were partialed out of the reported validity coefficients in all categories of test use (Jensen, 1980, chap. 8; Jensen, 1984). The validity of the single G-score of the General Aptitude Test Battery (GATB), for example, when averaged over 537 studies of 446 different occupations, is higher than the multifactor validity coefficient based on the multiple correlation between all nine of the GATB aptitudes and the job performance criteria, with the general factor partialed out (+.27 vs +.24). (Also recall that multiple correlations are always biased upward, whereas zero-order correlations are not.) Although *g* has predictive validity for performance in practically all jobs, a clerical speed and accuracy factor and a spatial visualization factor also add a significant increment to the predictive validity of the GATB for certain clerical and skilled blue-collar occupations. The average predictive validity coefficients of each of the nine GATB aptitude tests, in 300 different occupations, are correlated +.65 with the *g* loadings of these aptitude tests. The predictive validity of *g* generally increases with job complexity and is highest in those occupations involving the least automatization of performance demands and the greatest amount of specialized training, constant new learning, judgment, novel problem solving, and responsibility.

The g loadings of various psychometric tests are highly consistent across different racial populations when they share the same language and general cultural background. In 10 independent studies in which test batteries comprising anywhere from 6 to 25 different tests were administered to large representative samples of black and white Americans, and a *g* factor was extracted separately from the correlation matrices in the black and white samples, the coefficients of congruence between the *g* factors obtained in the black and white samples of the 10 studies ranged between +.993 and +.999, with a mean of +.996. Such congruence coefficients indicate virtual identity of the *g* factor in the black and white populations (Jensen, 1985; Naglieri & Jensen, in press).

Even the *g* loadings of the WISC subtests obtained in the population of Japan on the Japanese version of the WISC are highly similar to the subtests' *g* loadings in the American standardization sample for the WISC, showing congruence coefficients above +.97 (Jensen, 1983).

SOME COMMON MISUNDERSTANDINGS ABOUT HIGHLY *g*-LOADED TESTS

Total scores on tests labeled intelligence tests, IQ tests, general ability tests, cognitive abilities tests, general aptitude tests, scholastic aptitude tests, and other variants of these terms are all very highly *g* loaded. This class of high-*g* tests in particular has been subject to considerable popular prejudice in recent decades and has accrued a number of common misconceptions and misunderstandings which have gained currency even among some professional psychologists. The acquiescence to some of the prejudices and mistaken notions about such tests by many psychologists and even by some psychometricians and people in the testing industry probably reflects a defensive attitude in the face of the more blatant popular prejudices against tests. A defensive attitude about tests too often results in overstating the limitations of tests and belittling the significance of the individual differences they measure, probably in hopes of warding off the antitest prejudice that has prevailed in the popular media (Herrnstein, 1982; Snyderman & Rothman, 1986). Listed below are some of the more subtle of the various misunderstandings of this type that I have encountered rather frequently in the psychological literature. Each one is stated here in the form of a question.

Do intelligence tests measure some innate characteristic of individuals? It is often said that tests cannot measure innate, that is, genetically conditioned, traits in individuals. If this were true, of course, it would be both logically and empirically impossible for any test to show a heritability coefficient significantly greater than zero. The heritability (h^2) of a metric trait is defined as the proportion of its total *variance* (a measure of individual differences) in a sample of some population that is attributable to genetic factors. The total nongenetic variance that is not due to measurement error is $r_{xx} - h^2$, where r_{xx} is the reliability of the measurements. Innumerable studies have found the heritability of highly *g*-loaded tests to be substantial, with values of h^2 falling mostly in the range from .50 to .80. The correlational data on twins, adopted children, as well as many other kinship correlations, in addition to genetic phenomena such as inbreeding depression (which is discussed later in this paper) cannot be plausibly explained without reference to models of polygenic inheritance. This conclusion is really not in dispute among the majority of modern geneticists and specialists in behavioral genetics.

Analyzing the total variance into components attributable to various genetic and nongenetic sources is conceptually no different from the analysis of variance attributable to the effects of experimental manipulations, as is commonly done in experimental psychology, or from the analysis of variance into components or factors as in principal components and factor analysis, or from the analysis of test scores into true-score and error components, as in classical measurement theory. All of these components-of-variance models are conceptually the same. And in all of them an individual's score (or any kind of single measurement in the analyzed sample) can be expressed as a weighted sum of the various components. The simplest quantitative genetic model for an individual's phenotype (i.e., observed characteristic or obtained score) is $P = G + E$, where P, G, and E are deviations from their respective populations means; the letters stand for phenotypic (P), genotypic (G), and environmental (E) or other nongenetic values. (More complex partition of the G variance and the E variance [and their covariance] into various components [such as additive, dominance, and epistatic gene effects, genetic variance due to assortative mating, and common and specific environmental effects], as well as their interactions is possible.) It necessarily follows that if the heritability is significantly different from zero, the phenotypic measurements (scores) must to some degree reflect individual differences in genotypes. Given the individual's P (i.e., observed score deviation from the population mean), the individual's estimated genotypic deviation, G, is h^2P. The standard error of measurement of genotypes can be expressed in a form that is perfectly analogous to the standard error of measurement of any score. If the heritability is h^2, the standard error of measurement of the genotype will be $\sqrt{\sigma_p^2(1 - h^2)}$, where σ_p^2 is the total phenotypic variance. Just as we can *probabilistically* test the significance of the difference between the obtained scores of two individuals, by the same logic we could *probabilistically* test the significance of the difference between two individuals' estimated genotypic values. Although this argument is theoretically correct, there is no conceivable practical value in such *estimated* genotypic values for individuals, because *estimated* values are of necessity *perfectly correlated* with the obtained scores; estimated scores are just obtained scores that have been pushed by some constant fraction (i.e., $1 - h^2$) toward the overall mean of the total distribution of obtained scores, and consequently still maintain all the same essential statistical relationships to one another. So there is no useful advantage to estimated genotypic scores, and they would have the added disadvantage of the unreliability of h^2, which, like any other statistic, is subject to sampling error. It would be a quite different matter, of course, if we could measure genotypes directly. If we could, they would *not* be perfectly correlated with the phenotypic values. (The expected correlation between genotypic and phenotypic values is the square root of the heritability, or h.) But of course we cannot measure genotypes for intelligence directly, nor can we do so for *any* polygenic trait, physical or psychological. Hence, to single out only mental ability tests in this respect is to be grossly misleading. The important theoretical point to be made here is that the common notion that individual test scores do not reflect genetic factors is simply wrong. It is just as conceptually wrong as to say that an individual's height or weight or skin color does not reflect genetic factors, although these and certain other physical traits may have considerably higher heritability than most measurable psychological traits.

Are observed mean differences between populations in test scores phenotypic or genotypic? It is often stated that the average differences in ability test scores between different populations, such as races and social classes, are "*only* phenotypic" differences. Inclusion of the word "only" is what makes the statement completely misleading and thus scientifically wrong. By definition, test scores (or any other trait measurements) are phenotypic, and hence to say that mean differences in test scores are "*only* phenotypic" is quite meaningless, and if it conveys to anyone the impression that test scores are different from any other trait measurements in this respect, it is simply wrong. A phenotypic difference per se affords no basis for inferring the degree to which either genetic or environmental factors contribute to the difference. The fact that the kinds of analysis required properly to estimate the proportions of genetic and nongenetic variance between groups have not been undertaken only means that we do not know the extent of genetic and nongenetic influences on the observed or phenotypic group differences. Neither source of influence has been ruled out by any research to date, nor, as yet, has there been any scientifically worthy estimation of the relative influences of genetic and environmental factors to the variation in average test scores between different racial populations. All the various lines of relevant evidence available to us today on this question can do no more than increase or decrease the subjective plausibility that genetic factors are involved in population differences. A recent survey of 1020 experts in psychometrics and behavior genetics reported that 53% believe that genes and environment are both involved in the mean black–white IQ difference, compared to 17% who attribute the cause only to the environment, with the remaining 30% feeling there is insufficient evidence for any conclusions (Snyderman & Rothman, 1986). But all beliefs regarding this question, by experts or by anyone else, can at present represent no more than subjective statements of plausibility. As scientists we should find no satisfaction in this fact.

Are the average differences in g-loaded tests that are associated with socioeconomic status and occupational categories entirely attributable to differences in amount of schooling and in the types of knowledge sampled by the tests? That something besides schooling and the knowledge content of tests is involved in the observed average differences between SES and occupational groups is indicated by two main lines of evidence. Full siblings reared together in the same family, and thus having the same SES background, often end up as adults in different occupational and SES levels, and these differences are positively correlated with their IQ differences in childhood and adolescence; the same is true for differences in the occupational status of fathers and sons (Jensen, 1973, chap. 6; 1980, chap. 8). Also, measurements derived from the electrical potentials of the brain (the "average evoked potential"), which have been found to be correlated with psychometric *g*, also show significant mean differences between groups differing in occupational level (Schafer & Marcus, 1973). Thus, occupational differences are found on *g*-correlated measures even when such measures involve no scholastic or cultural content. Moreover, it is not true that occupational level is solely a function of amount of education; when amount of education is held constant, a part of the positive correlation between IQ and occupational status remains. IQ, or rather the *g* ability it measures, acts as a threshold variable with respect to educational attainments, with higher levels of IQ being probabilistically a necessary, but not sufficient, condition for passing higher educational hurdles. However, there also remains a correlation between individuals' SES of origin (i.e., the SES of their parents) and their adult educational and occupational status that is completely independent of *g*. It has been noted that if occupational status were completely dependent on *g* ability and not at all dependent on adult individuals' SES of origin, the present advantage of white middle-class children over working class children would be reduced by one-third, and the relationship between adult occupational status and *g* ability, or IQ, would be correspondingly increased (Humphreys, 1984, p. 240).

Is intelligence, or g, the same as general learning ability? The answer here depends on the sense in which the term *learning* is used. *Learning* has one general operational meaning under which are subsumed two importantly distinct meanings.

Learning, in the most general sense, may be defined operationally as any change in the probability of making a particular response to a particular stimulus, when the change in response probability is not attributable to fatigue, maturation, senility, sensorimotor impairment, brain damage, or drug effects. The two distinct meanings of "learning" may be described as (1) *comprehension* (i.e., grasping concepts, "getting the idea," "catching on") and (2) *improvement with practice.*

When people speak of school learning, they are referring mainly to learning in the sense of *comprehension*. Striking individual differences are observed in children's rates of advancement in acquiring new and progressively complex concepts. No sooner has a child in school "caught on" to one idea than he or she is confronted by a new one, which is probably at a higher level of complexity. The traditional "3 Rs" are essentially of this nature. Learning in this sense, as the acquisition of concepts and comprehension of new and progressively complex material, is factor analytically indistinguishable from general intelligence, or *g*. Highly *g*-loaded tests are good predictors of learning in this sense. This is true even when the *g*-loaded tests have no information content in common with the criterion measures of learning. The correlation is not a result of common elements in the two measures, themselves, but is a result of the fact that the two measures depend on the same brain processes. Hence, IQ tests, or any highly *g*-loaded tests under whatever label, are highly correlated with scholastic performance and with conceptual comprehension in any setting at any age.

Improvement in performance with practice is quite another story. In this case, either the necessary concepts have already been grasped or the material is so simple as to present no problem in terms of comprehension. Practice merely increases facility of performance or adds to the acquisition of highly similar information at the same level of conceptual complexity. What is commonly referred to as rote learning is largely of this nature. One can memorize a string of nonsense syllables or the multiplication tables by repetition, and speed and accuracy of recall will increase with practice. Learning in this sense, as improvement with practice, has relatively little correlation with *g*. Moreover, no one has yet been able to discover any general factor of learning ability in this category of learning. What little general factor there is among various measures of learning that represent merely improvement with practice is the same factor as psychometric *g*. But most of the variance in rate of improvement in performance with practice is quite narrowly *task specific* and does not reflect a *general* learning ability. Apparently there is no general learning ability independent of psychometric *g*. Even when original acquisition entails a good deal of conceptual comprehension, and hence is highly correlated with *g*, repeated practice tends to gradually *automatize* what has been learned. Automatization of learning conserves *g*, so to speak, and frees it for other, more novel, purposes. For example, learning to read music, for the beginning student, demands full attention, and rate of progress is quite *g* correlated. For an accomplished musician, however, reading music has become almost completely automatized. The musical score is seemingly transmitted automatically and directly to execution of the notes on the performer's instrument, and the performer can focus all his attention on intrinsic aspects of musical interpretation and expression. As I am writing this, the act of forming letters with my pen is completely automatic, yet when I was in the first grade in school, simple penmanship very likely demanded my total concentration. When measures of individual differences in automatized performance of any kind are factor analyzed along with a variety of psychometric tests, it is found that the automatized skills have relatively low *g* loadings, in quite marked contrast to their substantial correlation with *g* during the early stages of acquisition. This type of phenomenon has recently been interpreted within the framework of information-processing theory in terms of automatic and controlled processing and attention (Ackerman, 1986, in press; Shiffrin & Schneider, 1977).

Is g essential for the achievement of worldly "success"? It is a popular belief that while *g*, or "IQ," may be importantly related to scholastic performance, it has little importance in the actual race of life once people are out of school. Family influence, "connections," motivation, personality, character, and sheer luck are thought to outweigh intelligence as determiners of worldly success. All these factors undoubtedly play some part in what people generally mean by "success." Where does *g* come in?

If by "success" we mean attained occupational status and all its socioeconomic correlates, there is ample evidence that *g* is quite highly related to this complex of variables. The IQs of school age children are substantially correlated with their adult occupational level. Occupations, like tests, differ in their *g* demands, and persons who score low in *g*, relative to the population, have a low *probability* of succeeding in those occupations with high *g* demands relative to other occupations. This is due, in part, to the differing amounts of *g*-demanding educational requirements of various occupations and in part to the differing *g* demands of the occupations themselves. The ability represented by *g* acts as a probabilistic threshold for successful performance, a threshold that differs markedly for various occupations. That is, exceeding a certain threshold of ability is a necessary but not sufficient condition for succeeding in a given pursuit. The correlation between *g* and occupational level is between about +.50 and +.70. It is not higher than this for three main reasons: (1) Other traits, interests, and special talents independent of *g* are also correlated with occupational level and with degree of success within various occupations; (2) part of the variance in occupations is attributable to differences in background, opportunity, and unknown or chance factors that are independent of personal characteristics; and (3) most occupations accommodate various activities having a fairly wide range of *g* demands, but not so wide as to be within the capability of the whole population. It is a fact that the standard deviation of IQs in various occupations progressively shrinks as we move up the occupational scale and a decreasing proportion of the population can meet the *g* demands of the successively higher level occupations. Exceedingly few persons below the 75th percentile in *g* ever become physicians, for example, and even fewer become mathematicians and scientists. Although g cannot account for all of the variance in occupational level, it accounts for more than any other measurable sources of variance, independent of g, that we have been able to discover.

THE FACTOR ANALYTIC ARGUMENT FOR g

In subjecting a number of variables to factor analysis, there is no mathematically compelling reason for a solution which extracts a general factor. Once the number n of significant primary factors has been determined, the total common factor variance is also determined, and through rotation of the factor axes this variance may be allocated to n factors in an unlimited number of ways. The original correlations among all the variables, insofar as they reflect the n common factors, can be mathematically reconstituted identically by the n factors regardless of how the factors have been rotated. Their positions after rotation, and the pattern of factor loadings on the original variables, of course, determine the interpretation of the factors. The interpretations will differ as the factor axes are rotated into different positions. The positions of the factor axes can be likened to the lines of latitude and longitude on a globe or a map. It is quite arbitrary that the longitude lines on this grid are all made to pass through the north and south poles and the latitude lines are all made to be parallel to the equator. A grid with any other reference axes could serve equally well to specify the exact location of any point on the face of the earth. Locations specified in terms of one set of reference axes can be mathematically transformed to any other set of reference axes. The same thing is true in factor analysis, and *mathematically* any given set of factors that accounts for the common factor variance among all the variables is as good as any other set of factors that accounts for the same variance.

How, then, can one argue that certain factor structures may be scientifically preferable to other structures? The two main pillars of the argument are (1) *simplicity* (referred to by Thurstone, 1947, as the criterion of *simple structure*) and (2) the directness of relationship of factor structure to natural phenomena that are completely independent of the methodology of factor analysis.

Rotation of the primary (or first-order) factor axes to a position that approximates Thurstone's criterion of simple structure as closely as possible makes sense in terms of the clarity of description of the factors in terms of the various homogeneous tests that were entered into the analysis. If several verbal tests and several spatial tests are factor analyzed together, for example, and we are able to extract two significant factors, it makes good sense to rotate the factors in such a way that one factor has very large loadings on all the verbal tests and very small or zero loadings on all the spatial tests while the loadings of the tests on the other factor are just the opposite. We could then unequivocally label one factor verbal ability and the other spatial ability. Any other position of the factor axes would blur this picture; each factor then might show moderate loadings on both types of tests, rendering factor interpretation difficult or impossible in terms of our knowledge of the tests themselves. It is on such grounds that *simple structure* is the generally preferred criterion for the rotation of factor axes.

Orthogonal rotation means that all the factor axes are maintained at right angles to one another and the factors are therefore uncorrelated. But there is one thing about factor rotation that is *not* arbitrary but is simply imposed on the results by a fact of nature. It is the fact that in the domain of tests of human mental abilities, no matter how homogeneous the tests may be, it is impossible to achieve as good a fit to the criterion of simple structure with orthogonal (i.e., uncorrelated) factors as with *oblique* (i.e., correlated) factors. With a variety of highly homogeneous tests (i.e., tests with a single type of item) a very close approximation to perfect simple structure can be achieved by oblique rotation of the factors. But this means that the oblique factors themselves are correlated with one another, and their common variance can be partialed out and represented as a second-order factor. The residualized first-order factors are then orthogonalized, because their common variance is moved up into the second-order factor. If there is more than one second-order factor, the process is repeated, with extraction of a third-order factor. The single highest factor in this hierarchical structure is *g*. In the domain of ability tests, the emergence of the *g* factor is the inevitable consequence of following the criterion of simple structure to its logical conclusion. At the level of factor analysis, the rejection of *g* necessarily implies the rejection of simple structure. This is why orthogonal factor rotation, as is obtained by Kaiser's (1958) popular varimax program, is simply wrong in the abilities domain—orthogonal factors never approximate simple structure as closely as oblique factors. If we accept the logic of simple structure, we must extract oblique factors, and the correlated factors must then be factor analyzed. Thus, a hierarchical factor structure, with *g* at the apex (as shown in Fig. 1), is the necessary consequence of the simple structure criterion. At the level of factor analysis, any argument against *g* will have to begin with an argument against simple structure. So far, no compelling objection to simple structure has been made.

Within the framework of factor analysis, the extraction of a *g* factor has the virtue of being consistent with the observed fact of nature that all tests of ability show positive correlations with one another (i.e., *positive manifold*) and it can therefore be assumed that they all measure some one factor in common, whatever the ultimate nature of this factor may be. Any method of factor analysis which does not permit extraction of the *g* factor merely obscures this important natural phenomenon of positive manifold. An inevitable and empirically demonstrable consequence of positive manifold is the fact that, on average, overall ability differences *between* individuals in the population are greater than the differences among various abilities *within* individuals.

It still remains to be demonstrated, however, that the factors resulting from a hierarchical simple structure represent anything more than just the mathematical machinations of factor analysis as applied to ability tests. It is fairly easy to see the inferred abilities in the loadings of various tests on the primary factors, which can usually be described in terms of the nature of the particular tests with the highest loadings on each factor. The *g* factor, however, cannot be described in terms of particular tests. It is a much higher level of abstraction than the primary factors and therefore seems more remote from observable "reality." So the question naturally arises, does *g*, more than other factors, correspond to any real phenomena *outside* the realm of factor analysis? If it does not, then it may perhaps be justifiably viewed as merely an artifact of the factor analytic method.

We may seek the answer to this question by looking for correlates of psychometric *g* that lie outside both psychometrics and factor analysis. If significant and substantial correlations are found, and if the correlations are larger than the corresponding correlations with factors other than *g*, I believe we are justified in claiming that *g* is not merely a methodological artifact but represents a real aspect of nature. We can be accused of *reifying g* only if we fail to find that *g* corresponds to some reality outside the realm of psychometrics and factor analysis and still claim a reality for *g*. On the other hand, if *g* is found to be related to natural phenomena that are observed or measured independently of the means of deriving *g*, then *g* cannot be a reification of a methodological artifact but must be viewed as a natural phenomenon in its own right.

NONPSYCHOMETRIC CORRELATES OF *g*

The g loadings of various tests are directly related to the heritability of the tests. Heritability, h^2, is the proportion of variance in a trait that is attributable to genetic factors. A widely used method for estimating the heritability of trait measurements is based on a comparison of identical, or monozygotic (MZ), twins, who have all of their genetic inheritance in common, with fraternal, or dizygotic (DZ), twins, who have approximately only half of their genetic inheritance in common. In hereditary traits, MZ twins are, on average, more alike than DZ twins. In such a case, the within-pair variance for DZ twins (s^2_{WDZ}) will be greater than within-pair variance for MZ twins (s^2_{WMZ}). The variance ratio $F = s^2_{WDZ}/s^2_{WMZ}$ hence indicates the degree of genetic inheritance. (This F ratio is automatically corrected for attenuation, since the same error variance in the numerator and denominator cancels out.) A statistically significant

F ratio warrants rejection of the null hypothesis, i.e., that the heritability of the trait is zero.

These *F* ratios, based on sets of MZ and DZ twins, have been determined for 11 subtests of the Wechsler Adult Intelligence Scale (WAIS) in two independent studies (Block, 1968; Tambs, Sundet, & Magnus, 1984). The *F* ratios in the two studies range from 1.36 to 4.51, with a mean of 2.26; 18 of the 22 *F*s are significant beyond the .05 level.

When the 11 WAIS subtests are ranked in the order of their *F* ratios, and the subtests are also ranked in the order of their *g* loadings (based on the WAIS national standardization data), the rank order correlations between *F* ratios and *g* loadings are +.62 ($p < .05$) for the Block data and +.55 ($p < .05$) for the Tambs et al. data. Thus there is a relationship between size of the *g* loadings of WAIS subtests and the degree to which the subtests reflect genetic variance, and the heritability of the *g* variance is greater than the heritability of the non-*g* variance in the WAIS.

Tests' g loadings are related to the correlations of the tests between family members. Correlations between members of the same family reflect both their genetic relatedness and the effects of their sharing a common environment. Neither of these effects has any connection with either psychometrics or factor analysis. Yet the *g* factor loadings of 15 highly diverse cognitive tests have been found to be correlated with family correlations on these tests in a large sample (927 families) of white Americans (Nagoshi & Johnson, 1986). When the pattern of *g* loadings of the 15 tests is correlated with the pattern of familial correlations (all disattenuated) on the 15 tests, the following correlations are obtained:

Between spouses	+.90
Mother–daughter	+.76
Mother–son	+.69
Father–daughter	+.59
Father–son	+.55
Sister–sister	+.42
Brother–brother	+.33
Brother–sister	+.26

The high correlation (+.90) between the 15-test profile of *g* loadings and the profile of spouse correlations on the 15 tests indicates that assortative mating is based largely on *g*. Hence the effect of assortative mating on the genetic variance of abilities in the offspring generation will most strongly increase the genetic component of the *g* variance relative to the genetic component of other ability factors independent of *g*. For heritable traits, the effect of positive assortative mating (i.e., a positive correlation between spouses) is to increase the total genetic variance in the assortatively mating population. Virtually all of this effect of assortative mating consists of an increase in the genetic variance *between* families; the genetic variance *within* families (i.e., variance between full siblings) is scarcely affected by assortative mating (Jensen, 1978).

Tests' g loadings are related to the degree of inbreeding depression of test scores. Inbreeding depression is the diminution of a heritable trait in the offspring of genetically related parents as compared with the offspring of genetically unrelated parents. The higher the degree of kinship between the parents, the greater is the average degree of inbreeding depression in the offspring. However, inbreeding depression is observed only in those heritable traits which involve genetic dominance, that is, the phenotypic expression of the trait is enhanced by dominant alleles, while recessive alleles have the opposite effect. Traits that are fitness characters in the Darwinian sense, and hence have been subjected to natural selection in the course of evolution, increasingly develop genetic dominance over many generations. When evidence of genetic dominance is found for any polygenic trait, it is evidence that the trait has been subjected to directional selection in past generations. Inbreeding increases the degree of homozygosity (i.e., the proportion of paired alleles that are alike [dominant–dominant and recessive–recessive] relative to unlike pairs [dominant–recessive]), which diminishes the trait-enhancing potential of the dominant genes, resulting in the phenomenon known as inbreeding depression.

At least 12 independent studies have reported this genetically predictable effect of inbreeding on mental test scores (reviewed by Jensen, 1983; Agrawal, Sinha, & Jensen, 1984), and no studies have reported contrary findings. The effect of inbreeding depression on the IQs of children of first cousins, as compared with children of unrelated parents, is about one-third of a standard deviation (i.e., 5 IQ points) for Wechsler IQ and about one-half of a standard deviation on the Ravens Matrices, a more purely *g*-loaded test than the Wechsler (Agrawal et al., 1984).

As I have shown in detail elsewhere (Jensen, 1983), the degree of inbreeding depression on the various subtests of the Wechsler Intelligence Scale for Children (WISC) is directly related to the subtests' *g* loadings. The rank-order correlation between 11 WISC subtests' *g* loadings and their index of inbreeding depression (i.e., the difference between inbred and noninbred children) is about +.80. Varimax-rotated factor loadings show markedly smaller correlations with the index of inbreeding depression than do the *g*-factor loadings. These results indicate that psychometric *g* reflects a biological aspect of mental ability that acts as a fitness character which has been subjected to natural selection in the course of human evolution.

The degree to which various tests display the effect of hybrid vigor resulting from outbreeding is related to the tests' g loadings. Hybrid vigor, or heterosis, is just the opposite of inbreeding depression. Heterosis is an enhancement of the dominant trait. It results from outbreeding, that is, mating between individuals who are less closely related genetically, in terms of common ancestry, than the average degree of genetic similarity between mates in the general population. Heterosis, however, is less pronounced and therefore harder to measure, because, in human populations, the average degree of inbreeding is so small that deviations from the average in the direction of outbreeding cannot be very great. Hence the effects of inbreeding and of outbreeding cannot be symmetrically distributed around the extremely low average coefficient of inbreeding in the general population. Thus more extreme degrees of inbreeding are possible than of outbreeding, and consequently inbreeding depression can be a larger effect in terms of deviation from the population mean than heterosis can be.

Stimulated by my (Jensen, 1983) observation that the *g* loadings of various tests are directly related to the degree to which scores on the tests displayed inbreeding depression, Nagoshi and Johnson (1986) took the next logical step and tested the corollary hypothesis, namely, that tests' *g* loadings would also be related to the degree to which the scores on various tests displayed heterosis, or the genetically predictable effect of outbreeding. They looked at a large sample of children in Hawaii who were the offspring of matings between Americans of European descent and Americans of Japanese descent. This was the outbred group. The control groups were the offspring of parents who were both either European or Japanese Americans in Hawaii. The parent groups were matched on background factors known to be related to psychometric intelligence, such as education and socioeconomic status. Degree of heterosis was measured by the difference, in standardized scores, between the mean of the outbred and the mean of the two control groups on each of 15 highly varied cognitive tests. The 15 tests were also factor analyzed within each group to determine the tests' *g* loadings. The correlation between the 15 tests' *g* loadings and the measures of heterosis was +.44, in accord with the theoretical prediction. When *g* factor scores were computed for all subjects, it was found that the outbred group scored, on average, about one-fourth of a standard deviation *higher* than the mean of the two control groups. (The inbred offspring of first cousins, in contrast, would fall about one-third to one-half of a standard deviation *below* noninbred controls on *g* factor scores.)

Evoked electrical potentials of the brain are related to g. Certain features of the electrical potentials of the cerebral cortex evoked by a simple visual or auditory stimulus are correlated with IQ (for reviews, see Eysenck & Barrett, 1985; Haier, Robinson, Braden, & Williams, 1983). The subject, with an electrode attached to his scalp, merely sits in a reclining chair and simply hears a randomly spaced series of auditory "clicks," stimuli that cannot be regarded as "cognitive" or "intellectual" by any reasonable definition. The subject is not required to make any overt or voluntary responses during the session. But brain waves are recorded and averaged by computer over a given time-locked epoch marked by the occurrence of each auditory stimulus, yielding a highly distinctive waveform termed the *average evoked potential,* or AEP. The main features of interest are the average latency, intraindividual variability in latency, amplitude, and complexity of the AEP. (The average total excursion of the line marking the waveform of the evoked potential within a uniform epoch provides an objective measure of the complexity of an individual's AEP.) All of these measures have been found to be correlated with scores on various IQ tests.

The most interesting finding from our standpoint, however, is that the AEP is more closely related to *g* than to any other source of psychometric variance. Hence the *g* derived from the factor analysis of conventional psychometric tests is clearly related to an electrophysiological measure of brain activity.

Eysenck and Barrett (1985) measured both the complexity and intraindividual variance of the AEP waveform in 219 subjects and combined these features in a composite measure which they correlated with the WAIS Full Scale IQ in 219 subjects. Greater complexity and lesser variability are associated with higher IQ. Correlations were obtained between this AEP-derived measure and each of the 11 subtests of the WAIS. These correlations differed across the various subtests, and they also differed after correction for attenuation. Some subtests were clearly more highly correlated with the AEP than were others. But here is the interesting point: The rank-order correlation between the sizes of the subtests' correlations with the AEP and the sizes of the subtests' *g* loadings was +.93. In short, the various subtests are correlated with the AEP to the degree that they are loaded on *g*. When the AEP measure was factor analyzed among the WAIS subtests, it had a loading of +.77 on the *g* factor.

A similar study of the evoked potential was performed independently by Schafer (1985) in a sample of 52 adults of average to superior intelligence (WAIS Full Scale IQs of 98 to 142). But Schafer used a different measure, based on the observation that the amplitude of the evoked potential (EP) decreases with repeated trials. Schafer subtracted the average EP amplitude of the second block of 25 trials (25 auditory "clicks") from the average amplitude of the first block of 25 trials. This index of EP habituation, as Schafer termed it, was found to be correlated +.59 ($p < .01$) with the WAIS Full Scale IQ. (When corrected for the restricted range of IQ in this sample, the correlation rises to +.73.) Schafer also obtained correlations between the EP habituation index and each of the 11 subtests of the WAIS. Figure 2 shows these correlations plotted as a function of the subtests' *g* loadings (first principal factor). The Pearson correlation (r) is +.80 and the rank-order correlation (ρ) is +.77 between the subtests' *g* loadings and the size of their correlation with the EP habituation index.

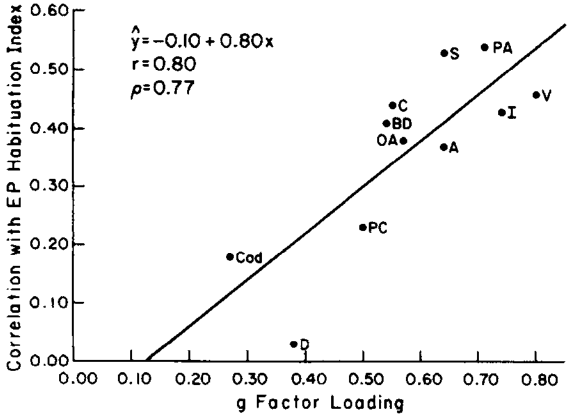

FIG. 2. Correlation of the habituation index of the evoked potential (EP) with Wechsler Adult Intelligence Scale (WAIS) subtests plotted as a function of the subtests' *g* loadings (i.e., first principal factor) in Schafer's study. WAIS subtests: I, Information; C, Comprehension; A, Arithmetic; S, Similarities; D, Digit Span; V, Vocabulary; Cod, Coding; PC, Picture Completion; BD, Block Design; PA, Picture Arrangement; OA, Object Assembly.

Moreover, Schafer found that no other factors, independent of *g*, that he could extract from the WAIS battery had any significant or appreciable correlation with the EP habituation index. The EP index reflects only the *g* factor.

The size of the mean black–white difference on various tests is directly related to the tests' g loadings. Sixty years ago, Spearman (1927) commented on his observation that the size of the mean black–white differences on a battery of 10 tests were "most marked in just those [tests] which are known to be most saturated with *g*" (p. 379). If Spearman's observation was confirmed with other test batteries in other samples of the black and white populations, it would constitute another example of the association of *g* with a variable outside the realm of factor analysis. Note that the *g* loadings of tests may be determined completely independently of the average black–white difference on the tests, so there can be no artifactual cause for a correlation between tests' *g* loadings and the magnitude of the black–white differences on the tests.

I have checked Spearman's original observation in 12 independent studies of large, representative samples of American blacks and whites that were administered anywhere from 6 to 25 diverse tests of cognitive abilities (Jensen, 1985; Naglieri & Jensen, in press). Spearman's observation is borne out in every study, and no study has been found which contradicts this finding. It indeed appears that the well-known average black–white difference on psychometric tests is much more a difference in *g* than in any other factor. In the largest and most representative samples the correlation between the mean black–white differences on various tests and the tests' *g* loadings is about +.80. This fact naturally has certain important and inescapable implications when tests are used for selection in education and employment, since the predictive validity of tests is largely attributable to their *g* loading. How the resulting "adverse impact" of testing on blacks should be dealt with is, of course, not a scientific question but a matter of public policy, on which opinions differ.

g is related to differences in reaction time in response to elementary cognitive tasks. Many studies in recent years have shown correlations between conventional highly *g*-loaded psychometric tests and reaction-time (RT) measures derived from a variety of elementary cognitive tasks. This literature is much too extensive to review in any detail here, so I will abstract a few of the general findings that are most germane to the present thesis. (For more general reviews, see Carroll, 1980; Eysenck, 1982; Vernon, in press).

An elementary cognitive task has been defined by Carroll (1980) as follows:

> An *elementary cognitive task* (ECT) is any one of a possibly very large set of tasks in which a person undertakes, or is assigned, a performance for which there is a specifiable class of "successful" or "correct" outcomes or end states which are to be attained through a relatively small number of mental processes or operations, and whose successful outcomes can differ depending upon the instructions given to, or the sets or plans adopted by, the person.

In general, ECTs are so simple that individual differences cannot be reliably measured in terms of number of right or wrong responses, as in ordinary psychometric tests, but must be measured in terms of response latency or reaction time (RT). ECTs usually involve little or nothing that could be called "intellectual" content or items of knowledge or skill that would not be possessed by any of the persons taking part in a study of an ECT. In my own studies of ECTs, for example, subjects, in order to qualify for participation, must demonstrate perfect scores on all of the information content of the ECT when the task is administered without time limit.

ECTs include simple and choice RT to simple visual or auditory stimuli (e.g., the onset of a light or a tone), visual scanning of a short series of digits for the presence or absence of a predetermined target digit, scanning of easily memorized series of 1 to 7 digits for the presence or absence of a designated probe digit, simple comparisons of letters or words as to whether they are the same or different with respect to physical, graphemic, or semantic characteristics. The extreme easiness of the ECTs used in my own research is shown by the fact that the median RT for Berkeley undergraduates on the most difficult tasks is less than 1 s and the response error rates are very low, averaging less than 5% of all responses.

Yet RTs derived from these very simple ECTs show significant and, in some cases, quite substantial correlations with scores on unspeeded psychometric tests. Even in the restricted range of ability in the college population, the correlations range between about −.10 (for simple RT) to about −.50 (for discrimination RT). The intertrial variability in RT (measured as the standard deviation of the subject's RTs over a given number of trials) is generally even more highly correlated (negatively) with psychometric test scores than is the average RT. The highest correlation found in our university sample so far is between an "oddman out" RT test and scores on the Advanced Ravens Progressive Matrices test given without time limit. The Ravens is a high-level test of nonverbal reasoning. In a sample of 71 students, the multiple correlation, based on median RT, median MT (movement time), and the *SD*s of RT and MT over 132 trials, was .60. This seems a remarkable correlation considering the simple nature of the "oddman" task. In a row of eight equidistant lights, a set of three lights goes on; one light (the "oddman") is always farther in distance from the other two lights. The subject's task is to turn off all three lights as quickly as he can simply by touching the "oddman" light. The subject first holds his finger on a central "home" button, then the three lights come on simultaneously, and the subject responds. RT is the time interval between stimulus onset and lifting the finger from the "home" button. Movement time (MT) is the interval between releasing the "home" button and touching the "oddman" button. The mean RT is only 460 ms; mean MT is 294 ms, and the response error rate is 2%—obviously a very simple task, not involving any knowledge or acquired skill, and too fast to allow what one would ordinarily think of as "cogitation." Yet it correlates .60 with scores on an unspeeded, difficult test of complex reasoning which is known to be one of the most highly *g* loaded of all psychometric tests. In another study, the "oddman" test showed a correlation of .62 with the WAIS Full Scale IQ (Eysenck & Frearson, in press).

Several generalizations concerning the relationship of RT in ECTs to psychometric *g* can be gleaned from this literature.

Probably the most important generalization has to do with *task complexity*. The best objective index of the complexity of an ECT is the average RT. For RTs up to as long as about 2 s, there is an increasing correlation between RT and *g*. In the one most direct study (Vernon & Jensen, 1984) of this phenomenon, the correlations of RT on each of eight tasks with *g* factor scores on the Armed Services Vocational Aptitude Battery were correlated −.98 with the complexity of the tasks (as indexed by the mean RT on each task). Also, groups that differ, on average, in *g*, such as retarded and normal vocational students and university students, gifted and average children, and blacks and whites, show average differences in RT, and the differences markedly increase as a function of

ECT complexity, even though all of the tasks are very simple for all subjects, with the longest average RTs of less than 2 s (Cohn, Carlson, & Jensen, 1985; Jensen, 1982, 1985; Vernon & Jensen, 1984). It is also possible to manipulate the *g* loadings of ECTs experimentally by creating "dual" or competing tasks, thereby increasing the demands on the individual's information-processing capacity (Fogarty & Stankov, 1982; Jensen, in press-a; Vernon, 1983).

Hemmelgarn and Kehle (1984) used the Hick RT paradigm (Jensen, in press-b), in which the subject's RT to either 1, 2, 4, or 8 light–button alternatives is measured. In this paradigm, RT increases as a linear function of the binary logarithm of the number of alternatives, a phenomenon known as Hick's law. The *slope* of this linear increase in RT may be viewed as a measure of the rate of information processing. This slope measure was correlated with scores on each of the 12 subtests of the Wechsler Intelligence Scale for Children—Revised (WISC-R) in a group of 59 elementary school pupils, with age partialed out of the correlations. The pattern of these 12 correlations had a rank-order correlation of $-.83$ ($p < .01$) with the pattern of the 12 subtests' *g* loadings. That is, the degree to which a WISC-R subtest is correlated with rate of information processing is highly related to the size of the subtest's *g* loading.

Vernon (1983) found a similar effect in a group of 100 university students who were given the WAIS and a battery of eight RT tasks. The multiple correlation was obtained between the eight RT tasks and each of the WAIS subtests. The pattern of these multiple correlations showed a correlation of .73 with the pattern of the subtests' *g* loadings. But the more important finding in Vernon's (1983) study was that just the *g* factor of the WAIS was correlated .41 with a composite score of all the RT tasks. The 11 subtests, with their *g* partialed out, showed a nonsignificant multiple *R* with the RT composite. In other words, virtually all of the correlation between the WAIS and the RT measures was attributable to the *g* factor of the WAIS.

The Wechsler tests were not devised with reference to factor analysis or *g* theory, and certainly they were never devised to produce a strong association between the subtests' *g* loadings and their degree of correlation with RT measures derived from ECTs.

As mentioned previously, the RT measures of various ECTs are themselves differentially correlated with *g* along some dimension of cognitive complexity of the ECTs. It may, therefore, seem rather dismaying to those of us who are over age 50 that speed of processing slowly declines beyond middle age, and the decline is greatest on the very same ECTs that are the most highly correlated with *g* (Ananda, 1985; Cerella, 1985; Cerella, Di Cara, Williams, & Bowles, 1986).

THEORIES OF *g*

All of the findings about *g* that I have reviewed in this paper, I believe, warrant a view of *g* as one of the major and fundamental variables in psychology, and a variable which unquestionably links psychology to biology and evolution. The nature of *g* itself, which is the basis of individual differences in what Lloyd Humphreys (1981) has aptly referred to as "*the* primary mental ability," is quite another story, which I have reviewed in considerable detail elsewhere (Jensen, in press-a). Suffice it to say here that at present there is no generally agreed-upon theory of the nature of *g* that goes much beyond the kinds of facts I have reviewed. The final explanation of *g* will depend upon future advances in our understanding of the physiology and biochemistry of the brain itself. The well-established and very substantial heritability of *g* measures leaves no doubt of the biological underpinning of individual differences in *g*, and it is in the province of the neurosciences that the nature of *g* will finally be understood. Without this direct approach, neurological theories of *g* will remain merely speculative. Cognitive psychology runs into great difficulty in providing an explanation of *g*, because the most elemental measurable components of cognitive processes are themselves correlated with one another, and the general factor extracted from these intercorrelated cognitive processes seems to be the same factor that we recognize as psychometric *g*. If so, we are virtually forced in the reductionist direction of seeking to understand this fundamental phenomenon at the level of brain physiology.

For the time being, what I think we can rather confidently say about *g*, in light of present evidence, is that *g* reflects some property or processes of the human brain that are manifested in many forms of adaptive behavior, and in which people differ, and that increase from birth to maturity, and decline in old age, and show physiological as well as behavioral correlates, and have a hereditary component, and have been subject to natural selection as a fitness character in the course of human evolution, and have important educational, occupational, economic, and social correlates in all industrialized societies, and have behavior correlates that accord with popular and commonsense notions of "intelligence."

REFERENCES

Ackerman, P. L. (1986). Individual differences in information processing: An investigation of intellectual abilities and task performance during practice. *Intelligence*, **10**, 101–139.

Ackerman, P. L. (in press). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin.*

Agrawal, N., Sinha, S. N., & Jensen, A. R. (1984). Effects of inbreeding on Raven Matrices. *Behavior Genetics*, **14**, 579–585.

Ananda, S. M. (1985). *Speed of information processing and psychometric abilities in later adulthood.* Unpublished doctoral dissertation, University of California, Berkeley.

Block, J. B. (1968). Hereditary components in the performance of twins on the WAIS. In S. G. Vandenberg (Ed.), *Progress in human behavior genetics*. Baltimore: The Johns Hopkins Univ. Press.

Carroll, J. B. (1980). *Individual difference relations in psychometric and experimental cognitive tasks*. Chapel Hill, NC: L. L. Thurstone Psychometric Laboratory, University of North Carolina.

Cerella, J. (1985). Information processing rates in the elderly. *Psychological Bulletin*, **98**, 67–83.

Cerella, J., Di Cara, R., Williams, D., & Bowles, N. (1986). Relations between information processing and intelligence in elderly adults. *Intelligence*, **10**, 75–91.

Cohn, S. J., Carlson, J. S., & Jensen, A. R. (1985). Speed of information processing in academically gifted youths. *Personality and Individual Differences*, **6**, 621–629.

Eysenck, H. J. (Ed.). (1982). *A model for intelligence*. Heidelberg: Springer-Verlag.

Eysenck, H. J., & Barrett, P. (1985). Psychophysiology and the measurement of intelligence. In C. R. Reynolds & V. Willson (Eds.), *Methodological and statistical advances in the study of individual differences*. New York: Plenum.

Eysenck, H. J., & Frearson, W. (in press). Intelligence, reaction time, and a new "Odd-Man-Out" RT paradigm. *Personality and Individual Differences*.

Fogarty, A., & Stankov, L. (1982). Competing tasks as an index of intelligence. *Personality and Individual Differences*, **3**, 407–422.

Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences*. London: Collins.

Garrett, H. E., Bryan, A. I., & Perl, R. E. (1935). The age factor in mental organization. *Archives of Psychology* (No. 176).

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**, 591–599.

Hemmelgarn, T. E., & Kehle, T. J. (1984). The relationship between reaction time and intelligence in children. *School Psychology International*, **5**, 77–84.

Herrnstein, R. J. (1982, August). IQ testing and the media. *Atlantic Monthly*, pp. 68–74.

Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and learning*. New York: Plenum.

Humphreys, L. G. (1984). General intelligence. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing*. New York: Plenum.

Jensen, A. R. (1973). *Educability and group differences*. New York: Harper & Row.

Jensen, A. R. (1978). Genetic and behavioral effects of nonrandom mating. In R. T. Osborne, C. E. Noble, & N. Weyl (Eds.), *Human variation: Biopsychology of age, race, and sex*. New York: Academic Press.

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

Jensen, A. R. (1982). Reaction time and psychometric *g*. In H. J. Eysenck (Ed.), *A model for intelligence*. Heidelberg: Springer-Verlag.

Jensen, A. R. (1983). Effects of inbreeding on mental-ability factors. *Personality and Individual Differences*, **4**, 71–87.

Jensen, A. R. (1984). Test validity: *g* versus the specificity doctrine. *Journal of Social and Biological Structures*, **7**, 93–118.

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

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

Jensen, A. R. (in press-b). Individual differences in the Hick paradigm. In P. A. Vernon (Ed.), *Speed of information-processing and intelligence*. Norwood, NJ: Ablex.

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,** 882–893.

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

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika,* **23,** 187–200.

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

Nagoshi, C. T., & Johnson, R. C. (1986). The ubiquity of *g*. *Personality and Individual Differences,* **7,** 201–207.

Schafer, E. W. P. (1985). Neural adaptability: A biological determinant of *g* factor intelligence. *The Behavioral and Brain Sciences,* **8,** 240–241.

Schafer, E. W. P., & Marcus, M. M. (1973). Self-stimulation alters human sensory brain responses. *Science,* **181,** 175–177.

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

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review,* **84,** 127–190.

Snyderman, M., & Rothman, S. (1986, Spring). Science, politics, and the IQ controversy. *The Public Interest* (No. 83), pp. 79–97.

Spearman, C. E. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology,* **15,** 201–293.

Spearman, C. E. (1927). *The abilities of man.* New York: Macmillan Co.

Tambs, K., Sundet, J. M., & Magnus, P. (1984). Heritability analysis of the WAIS subtests. A study of twins. *Intelligence,* **8,** 283–293.

Thorndike, R. L. (in press). Stability of factor loadings. *Personality and Individual Differences.*

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

Vernon, P. A. (1983). Speed of information processing and general intelligence. *Intelligence,* **7,** 53–70.

Vernon, P. A. (Ed.) (in press). *Speed of information-processing and intelligence.* Norwood, NJ: Ablex.

Vernon, P. A., & Jensen, A. R. (1984). Individual and group differences in intelligence and speed of information processing. *Personality and Individual Differences,* **5,** 411–423.

Wherry, R. J. (1959). Hierarchical factor solutions without rotation. *Psychometrika,* **24,** 45–51.

Received: July 1, 1986.