Ability Testing

COMMENTARY: VEHICLES OF *g*

Arthur R. Jensen
University of California, Berkeley

According to the Buros *Mental Measurements Yearbook,* there are some 120 different standardized tests of "intelligence" and "scholastic aptitude." The correlations among all these tests range from about .60 to .90, averaging approximately .75. If the matrix of correlations among all these diverse tests were subjected to a hierarchical factor analysis, what would be their correlations with the highest-order common factor? (Following Spearman, this factor is usually labeled *g,* for *general factor.*) If most of the tests have little variance in common besides *g,* one can confidently predict (from the formulation $r_{xy} = r_{xg} \times r_{yg}$) that the various tests would have factor loadings (i.e., correlation between test and factor) mostly in the range of .80 to .90. Some tests are correlated in part, of course, because they also have certain group factors in common, such as verbal, numerical, and spatial. But we know from countless factor analyses of tests of mental ability that one and the same *g* factor, discovered by Spearman in 1904, is the predominant component of variance in all measures of complex cognitive abilities, of which various composites constitute the standardized tests that are given labels such as intelligence (IQ), general aptitude, scholastic aptitude, vocational aptitude, learning potential, differential aptitudes, cognitive abilities, and assessment battery. Total scores on all such tests scarcely differ in their *g* loadings, and largely for this reason they typically rank order persons in much the same way, despite the tests' often vastly different appearance in information content and form of response.

This phenomenon was explicitly stated by Spearman (1927) as his theorem of *the indifference of the indicator.* In his words,

> This means that, for the purpose of indicating the amount of *g* possessed by a person, any test will do just as well as any other, provided only that its correlation with *g* is equally high. With this proviso, the most ridiculous "stunts" will measure the self-same *g* as will the highest exploits of logic or flights of imagination. . . . And here, it should be noticed, we come at last upon the secret of why all the current tests of "general intelligence" show high correlation with one another, as also with *g* itself. The reason lies, not in the theories inspiring these tests (which theories have been most confused), nor in any uniformity of construction (for this has often been wildly heterogeneous), but wholly and solely in the above shown "indifference of the indicator." Indeed, were it worth while, tests could be constructed which had the most grotesque appearance, and yet after all would correlate quite well with all the others. (pp. 197–198)

Spearman's statement is as true today as in 1927. Each new test of mental ability that comes on the market is claimed to be superior to all others, or to have certain added advantages, or to measure abilities or aspects of performance that are neglected by previous tests. In looking over new tests for which such claims are made, I consider it a safe bet that when they are factor analyzed in a matrix with a number of other mental ability tests, they too will be as highly *g* loaded as the others, bearing out Spearman's principle of the indifference of the indicator.

The principle of the indifference of the indicator is undoubtedly correct. Its interpretation, however, needs to be critically challenged. Although it is true that all cognitive tests are *vehicles* of *g,* whatever other sources of variance they may have, and it has proved impossible to devise a mental ability test that is not *g* loaded to some degree, the strictly formal psychometric features of the vehicle per se, quite aside from the inevitable *g,* are nevertheless of considerable importance. This importance should not be eclipsed by Spearman's principle. The importance of the psychometric properties of the vehicle in its own right justifies a variety of vehicles and continuing effort to devise new and better ones.

The crucial distinction between the vehicle and the factor(s) it carries divides the field of research in two: (a) the theory of mental ability, which focuses on the psychological and physiological nature of the factors (including *g*) found in psychometric tests and on the construct validity of the factors, aside from any particular test or its practical utility, and (b) psychometrics, which concerns the measurement properties of the particular vehicles of mental abilities and the validity of their practical uses. The adequacy and efficiency of the vehicles are determined according to multiple criteria, some more important than others for any particular purpose, hence justifying a variety of vehicles for measuring one and the same construct, much like the need for using different kinds of thermometers for measuring the temperature of a freezer, a living room, a person, a kitchen oven, and a blast furnace.

Indeed, since Spearman's time, the vehicles of *g* have improved considerably, and we know much more about their psychometric properties. In recent years, for example, a whole new field of psychometric technology has

Address correspondence to Arthur R. Jensen, School of Education, University of California, Berkeley, CA 94720.

developed out of concern with *test bias* when measurements are obtained in populations that are heterogeneous in social class, race, cultural background, and gender. There are multiple psychometric indicators of systematic bias, based on statistical analysis of specific subgroup differences in such important variables as reliability, construct and predictive validity, factor structure, rank order of item difficulty, item characteristic curves, the relative attractiveness among the distractors in multiple-choice items, the effects of coaching and practice, and the correlations of test scores with criteria outside the realm of mental tests. These psychometric methods are now being used to detect various kinds of bias in existing tests and to eliminate biases in the construction of new tests. Tests can be unbiased without damaging their *g* factor validity or their practical predictive validity. In fact, eliminating group biases generally improves test validity, and vice versa. Research of this kind affords the possibility of using tests that are unbiased with respect to different subpopulations (e.g., blacks vs. whites, males vs. females) for any of the legitimate practical uses of tests, such as the assessment of children with learning problems and selection based on the prediction of training outcomes and job performance.

Computerized testing, in conjunction with recent developments in item response theory, is another innovation. It greatly increases the efficiency of test administration by quickly zeroing in on just those items that are at the most suitable level of difficulty, and hence the most discriminating, for each person's level of ability. Yet the scaled scores represent the same latent trait (e.g., *g*) throughout their full range in the standardization population. This technique optimizes the trade-off between testing time and validity.

The ubiquity of Spearman's *g* in all manner of complex cognitive activity argues that some tests should be expressly designed to maximize their *g* construct validity, along with having optimal psychometric properties for their practical usefulness in particular situations and populations. This aim does not depreciate other kinds of psychometric instruments, of course, as other variables besides *g*, particularly in the personality domain, also contribute to assessment or prediction of complex performance outside the testing room. One of the most remarkable findings of recent research, however, concerns the striking degree to which *g* is the chief "active ingredient" in the practical predictive validity of tests. Partial out any test's *g* component, and its validity drops to near zero. A spatial visualization factor, independent of *g*, slightly enhances prediction of performance on certain jobs involving mechanical comprehension or spatial relations, and a clerical speed and accuracy factor slightly enhances validity for clerical jobs. But scarcely any other factors independent of *g* appreciably increase prediction of training outcomes or performance for the vast majority of jobs. Specific job knowledge tests are valid for the corresponding jobs, but this is partly attributable to their *g* loading; the specific knowledge variance reflects past achievement, rather than an aptitude, and its predictive validity, independent of *g*, reflects the narrow relevance of the specific knowledge to performing a particular job. Studies of the General Aptitude Test Battery and the Armed Services Vocational Aptitude Battery show that by far the single best predictor of a test's predictive validity (corrected for attenuation) for success in training and performance in a wide variety of jobs is the test's *g* loading (Jensen, 1980a, pp. 735–736; Ree, Earles, & Teachout, 1992). The mounting evidence for *validity generalization* largely reflects the fact that an extremely wide variety of occupations make *g*-loaded demands; hence, any highly *g*-loaded test has useful predictive validity for performance in a wide variety of occupations (Schmidt & Ones, 1992). Other factors add surprisingly little.

A corollary of this conclusion is that *subtest analysis*, or the interpretation of a person's pattern, or profile, of subtest scores (on batteries such as the Wechsler Intelligence Scales, the Kaufman Assessment Battery for Children, and the Armed Services Vocational Aptitude Battery), is highly suspect. Such profiles constitute a set of ipsative scores from which *g* has been largely (but not entirely) removed. Recent studies of the validity of ipsative score profiles and subtest analysis find them to be practically worthless (Glutting, McGrath, Kamphaus, & McDermott, 1992; McDermott, Fantuzzo, & Glutting, 1990; McDermott, Glutting, Watkins, & Baggaley, 1992; Ree & Earles, 1992; Ree, Earles, & Teachout, 1992). What scant validity subtest analysis may have is probably attributable in part to the slight *g* variance that ordinarily remains in ipsative scores, unless *g* is explicitly removed by regression.

Psychometric *g* appears to be one and the same factor whether it is derived from a matrix of *between-family* (BF) correlations or *within-family* (WF) correlations among a number of diverse tests. The rationale for comparing BF and WF correlations is explicated elsewhere (Jensen, 1980b). Briefly, a BF correlation of, say, tests *x* and *y* is the correlation between the mean of *x* scores and the mean of *y* scores obtained by the siblings in each family in a large sample of families. Hence, a BF correlation should reflect any cultural, socioeconomic, or other environmental differences between families that might systematically influence children's test performance. A WF correlation is based on the difference between siblings (reared together) on test *x* and on test *y*. Hence, a WF correlation does not reflect the effects of whatever variables are uniquely involved in test-score differences between families. When the *g* factor was extracted from the BF correlations and from the WF cor-

relations among 15 highly varied tests given to large samples of Americans of European ancestry (AEA) and Americans of Japanese ancestry (AJA), the congruence coefficient (an index of factor similarity scaled 0 to ±1) between the BF *g* and WF *g* was +.99 in both the AEA and the AJA samples (Nagoshi, Phillips, & Johnson, 1987). The authors concluded: "The similarity of the BF and WF [factor] structures suggests that the genetic and environmental influences underlying cognitive abilities are 'intrinsic' in nature, that is, not just due to between-family differences in culture, status, values, and fortuitous cross-assortative mating" (p. 305).

Using scores on a number of tests obtained from many sets of monozygotic and dizygotic twins, it is possible to analyze the correlations among tests into their genetic and environmental components and derive both genetic and environmental correlations among the tests. This was done with a battery of eight specific cognitive ability tests (SCA) involving verbal, spatial, speed, and memory abilities and three kinds of scholastic achievement (reading, math, and language) measured by the Metropolitan Achievement Tests (MAT). It was found that the cognitive tests and achievement tests are correlated with each other mainly by virtue of their genetic correlations, which average about +.60; the environmental correlations average about +.06 (Thompson, Detterman, & Plomin, 1991). Moreover, although it was not mentioned by these authors, the total matrix of genetic correlations among all of the SCA and MAT variables is accounted for by just a single, or general, factor. (That is, the correlation matrix is of unit rank, as proved by Spearman's *vanishing tetrads* criterion.) Could it be that *g* reflects scarcely anything besides the common genetic variance in mental tests? This intriguing possibility suggested by this study awaits replication with other sets of diverse tests.

It is a fact that the correlational structure of mental tests from which *g* and other factors are derived is much more stable across generations (in the same population) than are the mean scores on these tests. Flynn (1984) amassed evidence that the raw scores on standard IQ tests, such as the Stanford-Binet and Wechsler scales, show a secular trend equivalent to about 3 IQ points per decade over the past half-century. This trend is evident in many industrialized countries around the world. Yet there is no evidence of secular change in either the correlations among ability tests, or their predictive validity, or their correlations with external variables, or the average differences between races and social classes. It is as if in every year since about 1930 a small constant value was added to the raw scores of every person tested.

Based on the overall upward-trending mean, Flynn (1987) has argued that the raw scores on IQ tests are only a loose correlate of the construct of intelligence, or *g*. He claims that while tests can validly assess persons' relative standing on *g* when they all are tested within a limited time period, the tests cannot validly compare persons tested a decade or a generation apart. By analogy, one could validly measure a group of persons' relative heights by measuring the lengths of the shadows they cast at, say, 11:00 a.m. in a particular locality, but one could not validly claim that another group of persons whose shadows are measured with equal reliability at 4:00 p.m. are, on average, really taller than the persons whose shadows were measured at 11:00 a.m. A valid comparison would require a method of measurement that is invariant across time and location, or at least could precisely transform persons' shadow lengths obtained at different times and locations to one and the same scale, as the Fahrenheit and Celsius scales can be transformed to the Kelvin scale.

If Flynn is right—and his argument is not easily dismissed—it means that the scores obtained with our conventional tests, as the vehicles of *g*, are like measurements of the shadows in this height analogy. However, it has not yet been determined with any certainty whether the upward secular trend in mental test scores is attributable merely to the conventional vehicles of *g*, because of the shadowlike property of their measurements (resulting, say, from increasing test-wiseness in the population), or to a real change in the causal basis of *g* itself, perhaps because of improvements in nutrition, hygiene, and health care in the general population. (The short period of time over which test-score changes have occurred rules out a genetic explanation.) It is not a farfetched possibility that the change is due to nutrition, considering that average physical stature in industrialized countries has increased over the same time period and by approximately the same amount (in standard deviation units) as mental test scores (Lynn, 1990). At present, however, we really have no way to determine the cause of the secular change in the overall difficulty level of *g*-loaded tests. This is a major problem for psychometrics. Its solution is essential for researchers who would study population trends in mean intelligence level.

As the most important factor in tests of mental ability in terms of its ubiquity and relative size among all of the factors in psychometric tests, its correlations with neurophysiological variables and with the efficiency of information processes in elementary cognitive tasks, and its relation to educationally, occupationally, and socially important criteria, the empirical reality of *g* is hardly disputable. The vehicles of *g*, however, must continually stand scrutiny. They warrant the best efforts of psychometric science to make them as worthy as possible.

REFERENCES

Flynn, J.R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95,* 29–51.

Flynn, J.R. (1987). The ontology of intelligence. In J. Forge (Ed.), *Measurement, realism, and objectivity* (pp. 1–40). Norwell, MA: D. Reidel.

Glutting, J.J., McGrath, E.A., Kamphaus, R.W., & McDermott, P.A. (1992). Taxonomy and validity of subtest profiles on the Kaufman Assessment Battery for Children. *Journal of Special Education, 26,* 85–115.

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

Jensen, A.R. (1980b). Uses of sibling data in educational and psychological research. *American Educational Research Journal, 17,* 153–170.

Lynn, R. (1990). The role of nutrition in secular increases in intelligence. *Personality and Individual Differences, 11,* 273–285.

McDermott, P.A., Fantuzzo, J.W., & Glutting, J.J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment, 8,* 290–302.

McDermott, P.A., Glutting, J.J., Watkins, M.W., & Baggaley, A.R. (1992). Illusions of meaning in the ipsative assessment of children's ability. *Journal of Special Education, 25,* 504–526.

Nagoshi, C.T., Phillips, K., & Johnson, R.C. (1987). Between- versus within-family factor analyses of cognitive abilities. *Intelligence, 11,* 305–316.

Ree, M.J., & Earles, J.A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science, 1,* 86–89.

Ree, M.J., Earles, J.A., & Teachout, M.S. (1992). *General cognitive ability predicts job performance*. Brooks Air Force Base, TX: Armstrong Laboratory, Training Systems Research Division.

Schmidt, F.L., & Ones, D.S. (1992). Personnel selection. *Annual Review of Psychology, 43,* 627–670.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. London: Macmillan.

Thompson, L.A., Detterman, D.K., & Plomin, R. (1991). Associations between cognitive abilities and scholastic achievement: Genetic overlap but environmental differences. *Psychological Science, 2,* 158–165.

Ability Testing

Postscript on Ability Tests, Testing, and Public Policy

Intelligence testing is undeniably one of psychology's most important contributions to practical affairs, yet it may in a sense be a victim of its own success, and it certainly is a continuing focus of often-heated public debate.[1] The preceding articles by Carroll, Humphreys, and Jensen help separate science from dogma in this highly controversial area—making clear the conclusions that can and cannot legitimately be drawn from nearly a century of research on ability tests with respect to what they measure and what the measures can tell us about individual and group differences. Several firm propositions emerge:

1. Intelligence tests (i.e., ability tests heavily weighted on the *g* factor, as discussed by the preceding authors) are excellent predictors of performance in many domains, ranging from school to a wide variety of occupations.
2. Efforts to supplant definitions of intelligence in terms of test scores with definitions in terms of cognitive processes have yielded progress at a theoretical level but so far have not enhanced our ability to measure intelligence.
3. Intelligence test scores are significantly correlated with many factors relating to social, economic, and family background that also predict performance.
4. Ability tests that are more nearly independent of background factors (and less weighted on *g*) uniformly fail to be competitive with intelligence, or general ability, tests for predicting school and job performance. Thus, tests designed to measure the *g* factor continue to be strongly relied on for selection by employers, the military, and admission officials in schools at all levels.
5. Some of the background factors are associated with differing opportunities to learn material or to develop skills that are tapped by intelligence tests.
6. The use of the tests for job selection and admission to educational programs must tend to penalize individuals who have been handicapped with respect to opportunities to learn.

Thus, efficiency in selection and placement conflicts with considerations of justice. Should practice be based on the objective of maximizing productivity, or should productivity be sacrificed in the interest of fairness? Science cannot provide the answer, but scientists cannot afford to be aloof from an issue on which feelings in large segments of the public run high, some groups even arguing for the abolition of intelligence testing.

To allow research on intelligence to advance and to generate its long-term contributions to the public good, the use of tests in research must be unhindered. In return for freedom to conduct the research, however, scientists need to shoulder a heavy responsibility, not only for protecting the rights of the individuals tested, as is now routine in research though not yet routine in applications, but for developing an ethical code regarding the publication of research findings that bear on group differences in intelligence and other psychological characteristics—findings that often prove inflammatory when accounts spread outside scientific circles. Somehow a balance must be found between the need for free exchange of research results among scientists concerned with intelligence and the need to be sure that no segment of our society has reason to feel threatened by the research or its publication.

———W.K.E.

1. For an informative account of the evolution of public debate on testing, see the collection of essays edited by M.M. Sokal, *Psychological Testing in American Society, 1890–1930* (New Brunswick, NJ: Rutgers University Press, 1987; paperback edition, 1990).