Psychology, Public Policy, and Law
2000, Vol. 6, No. 1, 121–127

Copyright 2000 by the American Psychological Association, Inc.
1076-8971/00/$5.00 DOI: 10.1037//1076-8971.6.1.121

TESTING
The Dilemma of Group Differences

Arthur R. Jensen
University of California, Berkeley

A revival of test bashing has followed the growing public disillusionment with racial preferences as a means for abrogating the adverse impact of tests on certain groups in educational and occupational selection. This, added to the racial differences in *g* and the slight and ephemeral effects of Head Start and other more intensive interventions aimed at decreasing racial group differences in scholastic achievement, is the dilemma of group differences. In the face of the apparent failure of equal educational opportunity to make all groups in American society equal in scholastic performance or in the test scores used in selection for higher education, jobs, and the Armed Forces, psychological tests are again being blamed and scorned. This could be averted in the public's perception by emphasizing tests' face validity in test construction without lessening the latent traits that account for the tests' validity.

The new wave of opposition against the use of mental tests in educational and personnel selection really has nothing to do with psychometric science. It is a public, political, and popular media phenomenon entirely driven by the fact that identifiable groups in our society have different distributions of test scores, all of them more or less normally distributed but with different means and different standard deviations, and with the means of some groups differing by more than one standard deviation.

This well known fact, in and of itself, would be trivial were it not for the fact that the best professionally constructed tests are valid correlates and predictors of a person's actual performance in school, in college, in Armed Forces training programs, and in many jobs. Therefore tests are useful in selection, the need for which is inescapable whenever the number of applicants exceeds the number of openings or the costs of failure are prohibitive (e.g., pilots). Professionally constructed, standardized, objective tests have a better proven track record for valid meritocratic selection than any other single means that has been tried, including letters of recommendation, personal interviews, biographical information, essays, and portfolios. Because high school grades in English, history, math, science, and a foreign language may also reflect a student's level of academic motivation and effective study habits—variables that are not as well reflected by test scores alone—a combination of grade point average (GPA) and test scores is the best predictor of academic performance in college that psychometricians have come up with.

If a typical selective college decided to admit a random sample of youths from the general population and maintained its usual grading standards, the validity

Correspondence concerning this article should be addressed to Arthur R. Jensen, School of Education, University of California, Berkeley, California 94720-1670.

coefficient[1] of the composite high school grades and test scores for predicting course grades and graduation would probably exceed .70. Playing roulette with that much predictive power, one could break the bank at Monte Carlo within an hour. In actuality, however, because of self-selection by the applicants and the still further selection by the big-name colleges and universities, the range of talent among those finally accepted is, of course, greatly restricted, typically to the top quarter of the general population in scholastic aptitude, which considerably diminishes the validity coefficient when based only on students who gained admission. The resulting validity coefficients are typically between .30 and .50, depending on the degree of selection. The University of California, for example, is mandated to admit only applicants whose combined GPA and SAT scores are in the top 12.5% of each year's high school graduates.

When the approximately normal frequency distributions of test scores of distinguishable groups in the population fall in somewhat different (though greatly overlapping) regions of the scale, any selection cut-score on that scale will result in different proportions of each group that falls above (or below) the selection cut-off. Assuming two groups, each with a fairly normal distribution of scores and about the same standard deviation in both groups, if the group means differ by one standard deviation (*SD*), and if the cut-score were placed at the mean of the higher-scoring group, about 50% of that group would exceed that cut-score. But only about 16% of the lower-scoring group would exceed that same cut-score. If the selection cut-score is placed at one SD above the mean of the higher-scoring group, 16% of the group will be selected, while only about 3% of the lower-scoring group will qualify. The higher the cut-score, the greater will be the percentage discrepancy between the selection rates for the two groups. Hence, any test (or any other form of assessment) on which groups differ, if used in the selection process, will have some degree of adverse impact on the lower-scoring group.

Adverse impact is most striking for the largest racial minority in the United States, consisting of persons of African ancestry. In competition with Whites (and Asians) in any test-based selection procedure used in higher education and in employment for which applicants are selected only from a segment of the total distribution of test scores that lies above the general average for all applicants, proportionally fewer Blacks are selected. A consequence of selection generally is disproportionate representation of various racial and ethnic groups in the more desirable colleges, jobs, and training opportunities. The news media persistently report instances of this outcome but rarely inform the public about certain psychometric and statistical realities that are essential for understanding the problem. These are either unmentioned, obscured, denied, or badly misrepresented in public discussions. Many of the proposed solutions have ignored or

[1] The validity coefficient (r_{PC}) of a predictor (e.g., a test or other predictive index) tells us that if two groups differ, on average, by an amount Z_P on a test or other predictive index (where Z_P is the standardized score), they will differ by the amount $r_{PC} \times Z_C$ on the criterion, such as college GPA (where Z_C is the standardized measure of the criterion). The same prediction also applies to any two individuals. However, the error of prediction (i.e., the standard error of estimate) for a group mean is smaller (by a factor of $1/\sqrt{N}$, where N is the number of individuals in the group) than for an individual.

slighted the relevant facts that make the juxtaposition of meritocratic selection and group differences so problematic.

Proposed Remedies for Adverse Impact

The most obvious remedy is to not use tests in any way for selection. Assuming that selection is an inescapable necessity, what means should replace tests? A random lottery, of course, would predictably yield a lower performance level of the chosen applicants relative to that of a group selected on the basis of any information that has some validity, that is, a positive correlation with the selectees' later performance. Various indicators of past performance on highly similar criteria are generally good options, but each such indicator has some shortcoming that causes appropriate test scores to enhance predictive power when used in combination with it.

A group of individuals who are equated for amount of education and for school or college grades shows a narrower range of the latent traits that contribute to predictive validity than does a random group. However, they still range quite widely in ability because of differences in the selectiveness and grading standards of diverse schools and colleges, and the types of courses on which each individual's GPA is based. Nevertheless, because GPA and test scores are substantially correlated, the sole use of GPA for selection usually results in a highly similar ranking of applicants, and strict top-down selection still has almost as much adverse impact as test scores or even test scores and GPA combined (Crouse & Trusheim, 1988).

A whole class of proposals for reducing adverse impact would base selection on devices that we know would generally lower the overall predictive validity, such as making tests optional for each individual, basing selection on traits other than cognitive abilities, or using what are called "nuanced" assessments consisting of short essays, portfolios displaying examples of creativity or hobbies, and biographical inventories and interviews. Besides having undemonstrated validity or demonstrably inferior validity for predicting performance, all of these methods are enormously inefficient and costly as compared to either school grades or objective tests of cognitive abilities and academic achievements.

In addition, there have been attempts to take advantage of the efficiency of tests while at the same time reducing their adverse impact by composing tests of items that are specially selected so as to reduce group differences in the test scores. The only problem is that when tests are specially constructed to reduce group differences, they also have much lower validity within each of the groups. A test with equally high validity for selecting Blacks as for selecting Whites will consequently discriminate more between the Black and White groups than does a test with relatively low validity within each group. The reason is that the psychometric validity of the tests is color blind; that is, both the within-groups and the between-groups components of variance that contribute the most to the test's validity reflect equally the same latent traits, or factors, whatever their cause.

The one latent trait that has proven to be the *sine qua non* of tests' predictive validity for any and every kind of performance calling for mental ability (i.e., learning, memory, grasping concepts, reasoning, problem solving, etc.) is the so-called *g* factor. As the symbol for "general mental ability," it is widely known;

its nature is not. Even so, it lies at the heart of the whole problematic nexus involving the nature of group differences, the merits of meritocratic selection in a diverse society, the legitimacy of using tests, their adverse impact on certain groups, and its redress by group preferences in college admissions and employment. Some established facts about *g*, therefore, must be considered.

The Central Role of the *g* Factor

The *g* factor is a complex scientific construct whose meaning cannot be adequately conveyed by a simple definition. In the simplest terms it is usually defined operationally as the highest-order common factor in a hierarchical factor analysis of a large number of highly diverse mental tests or tasks, provided there is an objective standard for grading the participants' responses. It is *g* that, as researchers say, "accounts for" the fact that all such mental tests are positively correlated to some degree, and *g* accounts for a greater proportion of the variance, or individual differences, than any other single factor that can be identified in the correlations among any large collection of diverse mental tests given to a representative sample of the general population. I will be the first to admit that this "definition" of *g* is inadequate for those who want to understand the concept (and the spurious controversy surrounding it) in some depth, but it will have to do for present purposes. My recent book on *g* (Jensen, 1998a; see also the special issue of *Intelligence* edited by Gottfredson, 1997) gives it a quite thorough exposition and presents the considerable empirical evidence for the following facts about *g* that I think are most germane for understanding the problem nexus involving *g* that I have outlined above. These facts underscore the "dilemma" in my essay's title.

- The *g* factor, more than any other aspect of psychometric tests, is correlated with such nonpsychological, noneducational, and nonsociological variables as the *heritability* of test scores; the genetic phenomenon known as *inbreeding depression,* and its counterpart, *heterosis* or hybrid vigor; *brain size;* the brain's *glucose metabolic rate* under test conditions; the latency and amplitude of *evoked electrical brain potentials;* brain *neural conduction velocity;* and the brain intracellular pH level (i.e., alkalinity). All of these physical variables are related to psychometric *g*.
- Given a battery of language-appropriate tests and nonverbal tests, racial group differences in *g* are indistinguishable from individual differences in *g* within each group. This seems to me the key for understanding group differences in the *g* nexus: that whatever *g* is, average group differences in *g* are simply aggregated individual differences in *g,* so the composition of racial group differences and individual differences are of the same essential nature.
- The *g* factor is the chief "active ingredient" in tests' practical predictive validity. This is shown by the fact that when *g* is statistically partialled out of the correlation between the test and an external criterion, the correlation falls to near zero. It is also shown by the fact that when various tests are rank ordered by their *g* loadings (i.e., the degree to which they correlate with the *g* factor), the order is very nearly the same as the rank order of their validity coefficients. No other factor (independent of *g*) that can be extracted from a collection of diverse tests has higher predictive validity than does *g*. This does not mean that other psychometric factors, in addition to *g,* may not usefully increase a test's validity, but other factors rarely override the effect of *g,* and then only for some

specific performance criterion that calls for a special ability or talent (such as visual–spatial reasoning or musical talent), a particular knowledge domain, or certain skills.

- The size of the average Black–White (B–W) difference (expressed in standardized units) on various psychometric tests varies as a direct function of the tests' differing *g* loadings. In other words, the B–W difference on psychometric tests (and their many external correlates as well) is mainly a function of *g*, so that the more a test measures *g*, the more it discriminates between unselected groups of Blacks and Whites. No other feature of tests is as highly correlated with the variable size of the mean B–W difference on various tests. Nor are socioeconomic status or other social background factors as sharply predictive of B–W differences on a given test as is the *g* factor. Because *g* is the primary effective factor both in the practical validity of tests and in the magnitude of the B–W difference in unselected groups, the conjunction of these two effects is the unavoidable cause of adverse impact when *g*-loaded tests are used in selection. This is a more serious matter than if the tests were merely biased in cultural content: first, because it is not possible to rid cognitive tests of *g* and still have them remain valid for any practical purpose; and second, because individual differences in the level of *g* resist intentional change.
- No method of psychological or educational intervention has yet demonstrated reliably the power to make sizeable or enduring upward changes in children's IQ, or particularly their level of *g*, whatever their initial position on the scale. Although spontaneous changes in IQ may occur over the period from early childhood to maturity, their causes have not been identified. Small spontaneous changes, either up or down, are common; large changes are relatively rare, and they go downward as often as upward. However, psychologists and educators do not know of any means for raising the *g* level of children who are at risk for unsatisfactory scholastic performance. The largest authentic gains in *g* that have been induced experimentally in children by the most intensive and extensive means ever tried amount to, at most, about one third of a standard deviation (equivalent to 5 IQ points), and it is not yet known if this amount of gain will last to maturity. It seems unlikely, therefore, that population differences in *g* averaging one standard deviation or more can be overcome in the foreseeable future or will ever be overcome by environmental manipulations alone. The best we can do at present is to try and improve the conditions of learning while recognizing that the rate of cognitive learning itself is inescapably linked to individual differences in *g* (Jensen, 1998b).

Mitigating the Dilemma of Testing and Adverse Impact

The dilemma arising from tests and group differences in selection cannot be made to vanish, but there are practices that would probably allay popular antagonism toward tests. Racial preference for certain groups, a practice which in recent years has overshadowed the original concept of affirmative action, especially in college admissions, is now generally the institutional remedy for the adverse impact of tests (and GPA) in selection (Jensen, 1991). As tacit racial quotas based on double standards for selection are rapidly losing public favor, then what means of selection (where selection is necessary) might be acceptable to the public's perception of fairness, besides treating every applicant as an individual without considering group membership?

The answer, I would suggest, is to avoid, or at least minimize, the use of typical "intelligence" tests for selection purposes, however psychometrically excellent estimates of *g* they may provide. Instead, wherever possible, I would use selection criteria that assess relevant past performance (such as GPA in certain subjects, or work experience) and tests of relevant prerequisite scholastic or job-related knowledge and skills. In addition, it should be made publicly well-known precisely what the nature of these selection criteria are: They are assessments of the individual's present achievements in those areas that are most functionally germane to the given purpose of selection. It is much easier for the public to perceive a test's face validity than to understand its construct validity or its validity generalization. If a test's content is obviously what all applicants have had the opportunity to learn (e.g., those subjects in the high school curriculum that have been officially declared prerequisite for college admission), the public is unlikely to disapprove the use of such a test for selection into colleges that are unable to accommodate more than some fraction of all of the applicants. Emphasizing tests' face validity need not harm either construct validity or predictive validity. These features are more or less ensured by Spearman's principle of "the indifference of the indicator" along with the empirical fact of predictive validity generalization across many different performance criteria. When the applicant pool is homogeneous in amount of prior schooling, a comprehensive scholastic achievement test can reflect *g* as well as a conventional paper-and-pencil IQ test. It also typically contains a larger verbal ability component, which enhances its validity for predicting college performance more than would a relatively pure test of *g*. It is hardly an accident that the College Board's Scholastic Aptitude Test (SAT) and the American College Testing Program (ACT) both come close to satisfying these conditions. (The SAT and ACT are correlated about .90.) Employers, too, can determine the kinds of job-related knowledge and skills they wish their new employees to have, and selection tests can be composed of these elements.

In industry and in the Armed Forces, it is often necessary to employ personnel who have no prior experience in the specific job that they will be trained to do. In this case, the aim of selection is to recruit those applicants who are mostly likely to do well in training. For this purpose, a high degree of face validity is hardly suitable. The most effective tests are those that not only tap a broad variety of knowledge and cognitive skills but also contain highly *g*-loaded items that get at reasoning about things and call for little specific knowledge, but rather involve making comparisons, grasping relationships, inducing rules, deducing consequences, and the like—tests like the Raven Progressive Matrices, for example. Computerized test programs are being developed that efficiently zero-in on the subject's level in a wide range of brief cognitive tasks, and the composite score is an effective predictor of a person's ability to learn novel material.

There is no real escape from *g* whenever cognitive tests of any practical value, or any other objective criteria of cognitive performance, are used. This implies that there is also no escape from adverse impact, except by imposing different selection standards for different population groups.

However, note that adverse impact is a phenomenon wholly related to group differences. It need not be seen as a problem if selection were only thought of in terms of individual differences. Thus the "dilemma" referred to in the title of my essay really boils down to the kind of question that science is unable to answer.

It is the old question of whether group rights should predominate over individual rights. This is inherently not a scientific question at all, but a philosophical and ethical one. I have spelled out my opinion about it elsewhere (Jensen, 1991), to the effect that insistence on individual rights will, both in the short run and in the long run, provide the best assurance of whatever fairness for *all* persons lies within the power of human endeavor to achieve. The emphasis, I believe, should be on furthering equal opportunity and equal treatment for all persons, and let group outcomes become what they may, rather than eliminating adverse impact merely by having group rights trump individual rights. On this point, of course, the argument devolves wholly on philosophic principle and social consent. Scientists may legitimately formulate predictions about the probable outcomes of different public policies, but they are no better qualified philosophically or ethically than any other citizens to choose which policy should be empowered. In our system of government, such decisions rest with the citizens, their elected representatives, and ultimately with the courts.

References

Crouse, J., & Trusheim, D. (1988). *The case against the SAT.* Chicago: University of Chicago Press.

Gottfredson, L. S. (Ed.). (1997). Intelligence and social policy [Special issue]. *Intelligence, 24,* 1–320.

Jensen, A. R. (1991). Spearman's *g* and the problem of educational equality. *Oxford Review of Education, 17,* 169–187.

Jensen, A. R. (1998a). *The* g *factor.* Westport, CT: Praeger.

Jensen, A. R. (1998b). The *g* factor in the design of education. In R. J. Sternberg & W. M. Williams (Eds.), *Intelligence, instruction, and assessment* (pp. 111–131). Mahwah, NJ: Erlbaum.

Received December 16, 1997
Revision received January 19, 1998
Accepted January 19, 1998 ■