DOCUMENT RESUME

ED 393 9`8 TM 024 975

AUTHOR	Laosa, Luis M.
TITLE	Intelligence Testing and Social Policy.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-95-32
PUB DATE	Oct 95
NOTE	41p.; Version of a paper presented at the Annual Meeting of the American Psychological Association (103rd, New York, NY, August 11-15, 1995).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Beliefs; Ethnic Groups; *Genetics; Intelligence Quotient; *Intelligence Tests; *Policy Formation; Public Policy; *Racial Differences; Research Utilization; *Scientific Research; Socioeconomic Status; *Test Use; Values

ABSTRACT

There is a resurgence of scientific and public interest and controversy centering on four interrelated themes: (1) intelligence testing; (2) racial, ethnic, and socioeconomic differences in measured IQ; (3) genetic and environmental influences on abilities; and (4) the role of scientific research in social policy. Given the polarization of views, dispassionate discussion is needed to help inform the ongoing debate. The aim of this paper is to contribute to this discussion by helping to clarify certain issues, assumptions, and concerns that, because they are often misunderstood or ign^ ed, tend to obfuscate the debate. It is proposed that it is values, attitudes, and beliefs, and not rigorous rules of logic, that typically govern the process of drawing policy implications from scientific data. The study of values, attitudes, and beliefs has traditionally been the province of psychology, and it is reasonable to conclude that psychology, more so than any other single discipline, holds the key to a better understanding of connections between scientific reseaich and public policy. (Contains 59 references.) (Author/SLD)

RESEARCH REPORT

INTELLIGENCE TESTING AND SOCIAL POLICY

Luis M. Laosa

Educational Testing Service
Princeton, New Jersey
October 1995

Intelligence Testing and Social Policy

Luis M. Laosa

Educational Testing Service, Princeton, New Jersey

Intelligence Testing and Social Policy

Luis M. Laosa

Educational Testing Service, Princeton, New Jersey

Abstract

There is a resurgence of scientific and public interest and controversy centering on four interrelated themes: intelligence testing; racial, ethnic, and socioeconomic differences in measured IQ; genetic and environmental influences on abilities; and the role of scientific research in social policy. Given the polarization of views, dispassionate discussion is needed to help inform the ongoing debate. The aim of this paper is to contribute to this discussion by helping to clarify certain issues, assumptions, and concerns that, because they are often misunderstood or ignored, tend to obfuscate the debate.

Intelligence Testing and Social Policy

Luis M. Laosa

Educational Testing Service, Princeton, New Jersey

Introduction

The recent publication of The Bell Curve (Herrnstein & C. Murray, 1994), a highly controversial book, has rekindled intense and often bitter public debate on significant scientific and societal issues, including particularly the following interrelated themes: (a) questions concerning the relative influences of and different roles played by genetic and environmental factors in the development of human intelligence, (b) arguments regarding racial, ethnic, and socioeconomic group differences in measured intelligence, (c) concerns about possible bias in the tests employed to measure intelligence, and (d) questions concerning the implications of scientific research on these issues for education and public policy. I say rekindled because essentially the same issues have been debated--with similar levels of intensity and acrimony--in prior periods in the history of scientific psychology (see, for example, Laosa, 1977, 1984; Oakland & Laosa, 1977; and Wigdor & Garner, 1982, for reviews of earlier phases of this ongoing debate). Thus, like a refractory strain of retrovirus, the issues tend to remain latent and from time to time resurge brusquely onto the fore of public consciousness. At no time yet has there been a full resolution of the significant issues; that is to say, a broad and lasting consensus of opinion around them has yet to be achieved. Divided opinions are clearly discernible not only in the society at large but also within the scientific community--not surprisingly, since in many respects the latter is a social microcosm of the former.

Given the present state of the science and the polarization of public opinion, we are not likely to see--at least in the foreseeable future--widespread agreement on the basic issues. Nevertheless, dispassionate and informed discussion now can help dispel frequently misunderstood or ignored assumptions and clarify points that too often obfuscate the debate and may lead to potentially harmful interpretations and applications of the available scientific data. This paper is an attempt to contribute to this discussion; the aim is to help raise the level of discourse and increase common ground in order to prevent misuses of scientific knowledge for political ends.

Because intelligence level is typically measured by means of standardized IQ tests, the available scientific evidence bearing on questions regarding the influences of genetic versus environmental factors on the development of intellectual ability rests largely on scores derived from such tests and similar measures of general intelligence. Conceptions of this construct--which specifically denotes general cognitive functioning (g) as assessed in the psychometric tradition of a general factor derived from a battery of diverse cognitive ability tests--and how to measure it have changed remarkably little in the past 50 years (Carroll, 1982, 1993; Kaufman, 1994; Lubin, Larsen, & Matarazzo, 1984; Plomin & Petrill, 1995; Ree & Earles, 1991; Sattler, 1982, 1988, chapters 6-11; Terman & Merrill, 1973; Wechsler, 1958). At the heart of the heated controversy rekindled by the publication of The Bell Curve is the view--espoused by many of the book's critics--that, because of the lower average scores by members of particular racial, ethnic, and socioeconomic groups, incorrect inferences will be made as to the abilities of persons from these groups and hence their educational, occupational, and employment opportunities will be limited or even denied. Moreover, the use of standardized test scores as a basis for making judgments regarding racial, ethnic, and socioeconomic group differences in abilities--and, more significantly, in the capability to develop these abilities--is seen by many as indefensible in light of strong allegations that there are biases inherent in standardized ability tests, which unfairly penalize persons from backgrounds different from those of White, middle-class, native speakers of English. Thus, the question of test bias and the relevancy and use of standardized ability measures remain central concerns of the movements for civil rights, equity and fairness in educational, occupational, and employment opportunity, and social justice.

Explanations of Differences in Measured Intelligence

Questions of how to assess and interpret individual differences in human intellectual abilities have long been of central concern to psychologists and educators. Some violent polemics have centered on the issue of interpreting data on intelligence tests. Two traditional views on intelligence have persisted since prior to the turn of the century: assumptions of <u>fixed intelligence</u> and <u>predetermined development</u> (see, for example, Hunt, 1961; and Laosa, 1977, for historical overviews). These two assumptions often underlie the ideas that intelligence is an innate dimension of personal capacity and that it increases at a relatively fixed rate to a level in a range predetermined at birth. The notions of fixed intelligence and predetermined development clearly have potentially adverse effects on education, employment, and occupational policies and practices, since they encourage neglect of intellectual development. The argument is often made that because intelligence is predetermined, no amount of cultivation can significantly increase it. The assumption that intelligence and other personal characteristics are fixed can easily lead to an unwarranted emphasis on the matter of personal selection and a corresponding underemphasis in the areas of training and personal growth (Hunt, 1961; Laosa, 1977, 1984). Moreover, to the extent that an overemphasis on innate ability as a determinant of performance is a societal belief, it can function as a self-fulfilling prophecy (Bjork, 1994).

Entering the debate in the wake of the publication of The Bell Curve, an article ("statement") signed by a total of 52 persons, a number of whom are eminent experts[1] in the field of mental testing, was published in the Wall Street Journal (WSJ; December 13, 1994, op-ed page). Prompted by concerns that the publication of The Bell Curve has led "many commentators" to offer "opinions about human intelligence that misstate current scientific evidence," the WSJ article states a series of "conclusions regarded as mainstream among researchers on intelligence." Its stated aim is "to promote more reasoned discussion on the vexing phenomenon that the research has revealed in recent decades." Specifically, it is a 25-point summary and interpretation of research results gleaned from the vast extant literature of studies published over many years on the topic of intelligence testing bearing on issues of heredity, environmental influences, and differences among racial, ethnic, and socioeconomic groups. The article is titled "Mainstream Science on Intelligence," but whether the article in its entirety reflects the mainstream of current scientific opinion and research in the areas of intellectual abilities and mental measurement remains an open question. Thus, responding to the WSJ article, Donald T. Campbell (personal communication, May 18, 1995)[2]--an eminent expert not among the signers--correctly notes that although Point 22 of the article was worded to profess scientific ignorance regarding whether or not the observed racial and ethnic group differences in mean IQ scores are innate, overall the organization of the other points seems to many readers to strongly imply that these group differences have a large innate component. Thus Point 14 concludes that heredity plays a larger role than environment in creating individual differences in intelligence, while points 7, 8, 19, 20, 21, 23, and 24, which deal with racial and ethnic group differences, fail to point out that the heritability coefficients are not applicable to these between-group comparisons.[3] Thus Point 5 asserts that "Intelligence tests are not culturally biased against American blacks or other native-born, English-speaking peoples in the U.S. Rather, IQ scores predict equally accurately for all such Americans, regardless of race and social class," without pointing out, however, that this validity for predicting success in predominantly White, native-born English-speaking environments would be so were differences (individual or between-group or both) in the abilities involved totally the result of differences in opportunities to learn. In other words, as Campbell notes, the unbiased validity asserted does not apply to the IQ tests as measures of innate ability.

It is a long established fact, which no one disputes, that mean differences in scores on standardized IQ tests exist among racial, ethnic, and socioeconomic groups. It is also true that the score distributions around the respective group means overlap considerably. The disagreements center on the explanations (i.e., interpretations) of the causes of the observed between-group mean differences. It is generally agreed that the observed within-group individual differences in general intelligence reflect both genetic and environmental influences. Disagreements exist, however, regarding the relative importance of these two influences and the degree of malleability potential of inherited intellectual characteristics.

Is there a plausible alternative to the explanatory hypothesis that the observed group differences in mean intelligence test scores are the result of genetic differences in intelligence? Campbell answers this question in the affirmative, proposing that between-group environmental differences are large enough to explain group differences in mean IQ, an explanation "so plausible that speculations about genetic differences are not needed" (personal communication, May 18, 1995; see also Campbell & Frey, 1970). To illustrate the argument, consider the acquisition of vocabulary, which is a ubiquitous component of widely used tests of intelligence and by many experts considered to be one of the very best measures of general intelligence, or IQ.[4] It is obviously learned, and the resulting individual differences in vocabulary are generally considered to be the result of the learning environment and innate differences in learning ability. Hence, Campbell argues, the average differences in environments between Black and White American children in opportunity to learn the particular vocabulary (and other component subjects, i.e., knowledge and skill domains) employed in intelligence tests are such that they are adequate to explain Black-White differences in IQ test scores, without the need for positing genetic differences. To test this hypothesis, he suggests, consider measuring for each child the vocabulary level (e.g., as measured by the Wechsler subscale) of the naturally occurring spoken language of his or her caregivers, playmates, and later school mates, the hours adults have read to the child from particular types of children's books, etc. The vocabulary environment would be perhaps the easiest component to measure (for example, as Hart & Risley, 1995, do), but wide average Black-White differences in learning environments would almost certainly be found also on all the other component subjects of intelligence tests. These learning experiences would include, for example, opportunity for uninterrupted play with building blocks similar to those used in IQ test items as measures of nonverbal problem-solving ability, and in general, socioeconomic differences in the availability of "educational toys."

Campbell thus sensibly recommends--for future study and instrument development efforts--that we try to measure the opportunities to learn as precisely as we do the individual's cognitive performance. And for adequate communication about group differences to our fellow professionals and the public, "it would seem imperative that when we provide information on group differences on IQ, we accompany it by equally detailed data on an 'EIPQ,' an Environmental Intelligence Producing Quotient" (Campbell & Frey, 1970, pp. 456-457). In other words, looking at each component subject of intelligence and achievement measures, one should score the environment in terms of the degree to which it has produced similar experiences.

Addressing the statistical confounding of socioeconomic status with race and ethnicity that occurs in the U.S. population, Point 23 of the WSJ article states:

23. Racial-ethnic differences are somewhat smaller but still substantial for individuals from the same socioeconomic backgrounds. To illustrate, black students from prosperous families tend to score higher in IQ than blacks from poor families, but they score no higher, on average, than whites from poor families.

Campbell argues that this assertion (Point 23) reflects a largely invalid interpretation of the results of the available research. Specifically, he explains, efforts to compare "equally" educated or "equally" wealthy Black and White samples have failed to equate on intelligence test item learning environments. The problem of interpretation lies in ignoring statistical regression (to the mean) artifacts or error and unique reliable variance in covariates. In other words, when groups differ in means and one must therefore take one extreme of each group's distribution on the matching variable in order to find matching cases, there always occurs, on the average, undermatching. Thus, for instance, when Black and White individuals are matched (or regression adjusted), the group differences in the true scores on the covariate are only partly removed; that is to say, a residual latent true-score difference remains in the control measure. More obviously, the Black and White members of each matched pair may differ from each other in schooling quality, if not in length of schooling (Campbell, personal communication, May 18, 1995; Campbell & Boruch, 1975; Campbell & Erlebacher, 1970/1975; Cook & Campbell, 1979; see also Achen, 1986).

Cast in terms of measurement validity--more specifically, in terms of population generalizability, which is an element of construct validity--the challenge is to ascertain whether or not the construct measured by the control variable generalizes across the different groups (Laosa, 1990, 1991). When conducting comparative research, proper attention needs to be given to assessing the tenability of the assumption of population generalizability. The object is to ascertain whether or not the constructs generalize--i.e., have invariant meaning--across the groups one wishes to compare (Laosa, 1979/1989, 1990, 1991). The question of population generalizability applies both to covariates and dependent variables.

Bias in Test Use

The possibility of bias in the use of tests has received considerable attention from the general public, researchers and scholars in psychology and education, and the measurement profession (e.g., Cole & Moss, 1989). Indeed, concern about possible bias in the use of standardized tests has been a dominant theme for the past three decades, emanating largely from concerns about civil rights, equal opportunity in education and employment, fairness, and social justice for individuals and groups whose lives may be adversely affected by decisions made on the basis of their test performance (Laosa, 1977, 1984). As Cole and Moss (1989) note, the wide diversity of views of the many parties concerned with test bias--including test critics, the courts of law, researchers and scholars, professional organizations, legislatures, the mass media, advocacy groups, testing organizations--and the implicit and often emotional assumptions people make that lead them to view the same information in different ways add to the complexity and difficulty of the task of attempting to judge whether test bias is a reasonable explanation of test score differences.

Although there is no widespread agreement on the definition of bias, to argue--as Campbell does--that the assertion in Point 23 (quoted above) of the WSJ article is an invalid interpretation of the available research data is tantamount to concluding that the use of standardized IQ tests for purposes of drawing inferences about group differences in general intelligence (as opposed to inferences about group differences in opportunities to learn) constitutes a biased use of these tests. Although the term bias has been used in a variety of different ways, technical definitions tend to be closely tied to validity theory (e.g., Cole & Moss, 1989; Messick, 1989, 1995). Thus, Cole and Moss (1989, p. 205) proffer the following definition: "An inference is biased when it is not equally valid for different groups. Bias is present when a test score has meanings or implications for a relevant, definable subgroup of test-takers that are different from the meanings or implications for the remainder of the test takers. Thus, bias is differential validity of a given interpretation of a test score for any definable, relevant group of test-takers." In other words, from the standpoint of a unified conception of validity (Messick, 1989, 1995), bias is a matter of population generalizability.

Although the definition of bias is thus simply stated, the determination of bias, like other determinants of validity, is a complex process. Because bias resides in the interpretation of a test score, not in the test per se, several interpretations might need to be considered for each test (Cole & Moss, 1989; Messick, 1989). While many measurement theorists emphasize the importance of focusing on the appropriateness of a use, they have seldom examined the way in which the use influences the meaning constructed for the test score. Noting this omission, Cole and Moss (1989) emphasize the importance of the context in which a test is used. If one ignores this context, they point out, it is impossible to choose among the many possible validity questions. Implicit in the concept of construct validity is the idea that the particular construct under consideration (e g., general intelligence) is hypothesized as a possible explanation of the scores on a test. There always are, however, other plausible hypotheses about the meaning of the test score besides the intended one. Yet, the application of validation theory in common practice tends to focus almost exclusively on the particular construct hypothesis and the evidence to support its plausibility (Cole & Moss, 1989). Such a focus overlooks the useful information that can be gained from considering rival hypotheses and the evidence that supports or refutes them (Cole & Moss, 1989; Cook & Campbell, 1979; Laosa, 1979/1989, 1990, 1991; Messick, 1989, 1995).

In its proper role, validation is guided by the generation of rival hypotheses or possible explanations for test scores, in addition to the construct hypothesis. Such hypotheses guide the search and bring forward the need for logical and empirical evidence. Indeed, it is the exploration of <u>rival</u> hypotheses that provides evidence about bias (Cole & Moss, 1989). Although there are many ways of accumulating evidence to support an inference, these ways are essentially the methods of science; however, as Messick (1989, p. 14) notes, validation "is not hypothesis testing in isolation but, rather, theory testing more broadly because the source, meaning, and import of score-based hypotheses derive from the interpretive theories of score meaning in which these hypotheses are rooted." In this way, research scientists and scholars can contribute significantly to the public debate. The hope is that whenever judgments, choices, and decisions are made--by the public, professionals, and policy makers--on the basis of test scores, they be made with informed discernment and fairness.

Nearly a quarter-century ago, Thorndike (1971) noted the increased questioning of the fairness of using standardized ability tests for certain racial and ethnic minority groups. He observed, as Linn (1989) reminds us, a lack of clarity in the definition of fairness and a shortage of evidence relevant to the question. Because of the substantial amount of work being undertaken at that time, however, Thorndike was optimistic that "clarification of concepts and expansion of the data base from which conclusions may be drawn can be expected in the near future" (p. 12). Considerable research has indeed been conducted during the past 25 years, and much has been written about item bias, bias in test use, and concepts of fairness. As Cole and Moss (1989) clearly demonstrate, however, clarity in definitions and evidence regarding the comparability of prediction systems cannot be expected to resolve the underlying value conflicts. To illustrate this point, Linn draws from the literature on college-admissions tests, which, although not intended as intelligence tests, have been the focus of considerable research helpfully relevant to tests generally. Thus, knowledge that the regression of first-year grades in college on high school grades and college-admissions test scores is essentially the same for Black and White students "is relevant to the decision of whether or not the test information should be used in the admissions process for black students. But," Linn enjoins, "a variety of other types of information regarding the likely consequences of using the prediction information in various ways is also needed" (p. 6). As Messick (1989, 1995) argues, test use needs to have both an evidential basis and a consequential basis. This need to analyze the consequences of test uses poses serious challenges for test users and test producers.[5] Whereas analysis of the likelihood of some consequences (e.g., short-term adverse impact) can be accomplished with relative ease--even though there may be differences in perspectives on the manner of analysis--it is far more difficult to evaluate consequences of a longer-term or more global nature. Linn offers as example the challenges involved in evaluating the consequences of the decision taken by the National Collegiate Athletic Association to require that the combination of grade-point average in core subjects and scores on college admissions tests exceed a specified minimum in order for the athlete to be eligible to compete during the freshman year of college. Studies were undertaken to investigate a variety of issues related to the policy as it was initially proposed, as well as to several alternative policies. These analyses addressed issues of likely adverse impact, differential prediction of grades, academic progress, and graduation. As Linn observes, however, many other issues considered relevant by supporters and opponents of the policy were not, and possibly could not have been, addressed--including, for example, the effects of the policy on the decisions of minority student athletes to take different courses in high school, on the guidance and support services (including test preparation courses) provided by high schools, on the likelihood that students who are not eligible their freshman year will still attend college, or on the actions of colleges to support athletes who are not eligible their freshman year, and the long-term effects on the education and employment of racial and ethnic minorities. The point of this case illustration "is not . . . to suggest that all these consequences should have been investigated before any action was taken, or even to suggest that they are all part of a complete analysis of bias in test use and interpretation. Rather, it is intended to show that judgments about what is a desirable and fair use of a test depend on a host of considerations and on the values that are attached to various effects" (Linn, 1989, p. 6). A measurement specialist could--as do the signers of the WSJ article (Point 5)--appropriately define the absence of predictive bias in accord with the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1985, p. 12) by finding that "the predictive relationship of two groups being compared can be adequately described by a common algorithm (e.g., regression line)."[6] It should be recognized, however, as Linn (p. 6) again enjoins, that "this definition neither corresponds to the meaning of the critic who charges test bias nor resolves the issue of how the scores of minority test takers should be used or interpreted." This point is recognized also in the Standards (p. 13), which acknowledges that its definition of predictive--or selection--bias is adopted "with the understanding that it does not resolve the larger issues of fairness." There is no doubt that concerns about the fair uses of tests for ethnic and racial minorities will continue to be a major theme in years to come. These concerns, moreover, are expanding to include--as Linn (1989) predicted--issues related to the testing of persons whose native language is different from English and of persons with disabilities.

Expert Opinion on Intelligence

Raised earlier was the question of what the "mainstream" of expert opinion is in the specialty areas of intellectual abilities and mental measurement. It is informative to consider the results of an opinion survey of psychologists and education specialists with expertise in areas related to intelligence testing. The survey, conducted by Snyderman and Rothman (1987), included several items relevant to the issues addressed in this paper. Thus, respondents were asked to indicate (a) each behavioral descriptor they believe to be an important element of intelligence and (b) whether or not they believe that the particular behavioral descriptor is adequately measured by the most commonly used intelligence tests. Over 95% of the respondents believe that abstract thinking or reasoning, problem-solving ability, and capacity to acquire knowledge are important elements of intelligence; yet, 20%, 27%, and 42% of these respondents believe that these three elements, respectively, are <u>not</u> adequately measured by the most commonly used intelligence tests. Next in descending order of rated importance, 70% to 80% of the respondents believe that memory, adaptation to one's environment, mental speed, and linguistic competence are important elements of intelligence; yet, 13%, 75%, 13%, and 14% of these respondents believe that these four elements, respectively, are <u>not</u> adequately measured by the intelligence tests. Finally, 68%, 62%, and 19% of the respondents, respectively, believe mathematical competence, general knowledge, and achievement motivation are important elements of intelligence; yet, 12%, 11%, and 72% of these respondents believe these three elements, respectively, are not adequately measured by the most commonly used intelligence tests.

Next, Snyderman and Rothman asked the respondents to indicate the extent to which the most commonly used intelligence tests are biased against Black Americans. Bias was defined as an average Black American's test score underrepresenting his or her actual level of the abilities the test purports to measure, relative to the average ability level of members of other racial and ethnic groups. The responses were given on a 4-point rating scale (1 = not at all or insignificantly biased, 2 = somewhat biased, 3 = moderately biased, and 4 = extremely biased). The reported mean rating for this question is 2.12 (SD = 0.8), showing that the respondents on the average believe there is some significant racial bias in intelligence tests. The survey also included a question identical to that on racial bias, except that it asks instead about bias against people of low socioeconomic status. The reported mean rating for socioeconomic bias is practically identical to that obtained for racial bias, at 2.24 (SD = 0.8).

The survey also asked respondents to express their opinion on the role of genetic differences in the Black-White IQ difference. Snyderman and Rothman report that 45% believe the difference is a product of both genetic and environmental influences; 15% believe the difference is entirely the result of environmental variation; and 24% do not believe there are sufficient data to support any reasonable opinion (14% did not respond to the question). Finally, the survey asked respondents their opinion on the role of genetic differences in socioeconomic status differences. The report shows that 55% believe the differences are a product of both genetic and environmental influences; 12% believe the differences are entirely environmental; and 18% feel there are insufficient data (15% did not respond to the question).

The next section of the paper focuses on emerging transformations in the manner in which researchers investigate genetic influences; these developments are bound to become part of the ongoing debate concerning differences in measured intelligence.

Molecular Genetics

Rapid advances in the field of molecular genetics during the past decade have opened a new era for genetic research, thus forecasting a new chapter in the history of research on genetics and human intelligence. These scientific advances in molecular genetics--i.e., the study of the molecular structure of the genes and the mechanisms by which genes control the activities of the cell--some researchers (e.g., J. C. Murray et al., 1994; Plomin, 1995; Plomin & Petrill, 1995) believe, will make it possible to identify specific genes responsible for hereditary influences on intelligence. Thus far, a number of genes have been identified that affect cognitive abilities. Most of these findings, however, involve rare disorders for which a single gene appears to be the necessary and sufficient cause of distinct types of mental retardation--such as PKU (phenylketonuria),[7] for example. Compared with single-gene disorders, it is believed that the task will be much more difficult for complex constructs such as general intelligence, which appear to be influenced by multiple genes as well as multiple environmental factors (Lander & Schork, 1994; Risch & Zhang, 1995). Plomin (1995) believes the challenge is to use the thousands of newly identified DNA (deoxyribonucleic acid) markers[8] in order to identify not the gene for intelligence, but rather the many genes, each of which makes a small contribution to the variance in the population. It is believed that, unlike Mendel's smooth and wrinkled seeds,[9] most behavioral dimensions and disorders are not distributed in simple either/or dichotomies. Such genes of varying effect sizes that contribute to common, complex quantitative traits are called quantitative trait loci (QTL); they are thought to "contribute interchangeably and additively as probabilistic propensities [rather than predetermined programming, so that] any particular QTL within a multiple gene system is neither necessary nor sufficient" (Plomin & Petrill, 1995, p. 22; see also Lander & Schork, 1994; Plomin, 1995; Risch & Zhang, 1995).[10] Significantly, evidence suggesting a QTL on chromosome 6 for reading disability has recently been reported (Cardon, Smith, Fulker, Kimberling, Pennington, & DeFries, 1994). Methodologically, the search for the intelligence genes--which has already begun (e.g., Plomin & Petrill, 1995)--will revolutionize genetic research on intelligence by recasting twin and adoption studies in terms of DNA correlations with IQ test scores.

The new opportunities and challenges for biological research also create an urgency to face the parallel challenges in the areas of social policy, law, and ethics. The ability of scientists to distinguish individuals genetically for forensic purposes, identify genetic predispositions for common and rare inherited disorders, and to characterize, if present, the genetic components of normal trait variability such as for height, intelligence, sexual preference,[11] or personality type has never been greater (Lander & Schork, 1994; J. C. Murray et al., 1994). Research seeking to identify the genes responsible for intelligent behavior will doubtless raise new and significant questions and controversies, as diverse interpretations, inferences, and hence policy implications will be drawn from the resulting data. At the same time, the new technologies may be applied in ways that amplify and reinforce old convictions and support existing institutional practices. For single-gene disorders, identification of the responsible genes has already led to serious concerns such as the potential for discriminatory and stigmatizing effects of access to genetic information by insurance providers and employers, and concerns about quality control and oversight mechanisms for genetic testing services (Knoppers & Chadwick, 1994; Nelkin & Tancredi, 1989). Indeed, one can already discern the beginnings of a new chapter in the ongoing debate on intelligence testing and social policy.

Values and Policy Implications

The final point of the WSJ article states:

> 25. The research findings neither dictate nor preclude any particular social policy, because they can never determine our goals. They can, however, help us estimate the likely success and side-effects of pursuing those goals via different means.

This point is partly well taken in the sense that public policies seldom emanate directly and logically from scientific research findings. Policy making is a highly political process in which the jostling of values, special interests, public opinion, attitudes, political expediency, ideology, precedence, and compromise are the factors that typically determine the final shape of a policy. A policy proposal may be written in such a way as to reflect faithfully the implications one draws from scientific data. The task of developing a proposal, however, is far from (a) whether or not it eventually becomes policy and, if it does, (b) whether or not it will, after surviving the policy making process, look anything like the original intent. Even farther away is the manner in which the policy is actually implemented. Point 25 of the WSJ article is well taken also in the sense that a given scientific finding may point to (and often does) more than one plausible interpretation and may thus suggest more than one policy implication. Thus Scarr (1994-1995)--a signer of the article--argues for Point 25, maintaining that one could argue that knowing that low IQ is heritable calls for more, not less support for those so disadvantaged through no fault of their own. On the other hand, it is also true that a particular interpretation of a given scientific finding may more likely suggest a particular policy direction than does an equally plausible alternative interpretation of the same finding.

Campbell thus objects to Point 25 of the WSJ article, arguing that the decision to interpret the available data as supportive of the view that social class and racial or ethnic group differences in intelligence are innate--and that therefore efforts to reduce them will be ineffective--will lead to policies that increase differences. He notes that the policy implication of this decision is to discontinue compensatory educational efforts. To illustrate his point, Campbell reminds us that on the basis of research findings by A. R. Jensen--one of the signers of the WSJ article--emanated the recommendation to establish separate curricula for Black and White students: rote learning for one, conceptual problem solving for the other (Jensen, 1972). If this recommendation were implemented, Campbell warns us, the quality of schooling for Black children, which is on the average already worse than that White children receive, would become even worse.

Logic and Psychology

A democratic nation's public policies are likely a fair reflection of its population's dominant values, attitudes, and beliefs. For this reason, we should not always expect the process of drawing policy implications from scientific research findings to stand up to the rigors of logical analysis. Indeed, one of the most profound and vexing problems in moral philosophy is that of providing a justification for value judgments.

The study of logic shows us that, as a mode of argument, deduction has a severe limitation: The content of the conclusion of a valid (i.e., sound) deductive argument is present in the premises. Invalid arguments that look something like valid deductions--and therefore can be easily confused with them--are called invalid deductions or deductive fallacies (Salmon, 1984). (Analogous observations can be made about mistakes in inductive reasoning; see, e.g., Salmon, 1984.) The British philosopher David Hume (1711-1776) saw clearly that value judgments cannot be justified by deducing them from statements of fact alone. Thus he wrote:

> In every system of morality which I have hitherto met with, I have always remarked that the author proceeds for some time in the ordinary way of reasoning, and establishes the being of a god, or makes observations

> concerning human affairs; when of a sudden I am surprised to find that instead of the usual copulations of proposition is and is not, I meet with no proposition that is not connected with an ought or an ought not. This change is imperceptible, but is, however, of the last consequence. For as this ought or ought not expresses some new relation or affirmation, it is necessary that it should be observed and explained; and at the same time that a reason should be given for what seems altogether inconceivable, how this new relation can be a deduction from others which are entirely different from it. (A Treatise of Human Nature [1739-1749], Book 3, Part 1, Sec. 1, cited in Salmon, 1984, p. 17)

In contrast with the social sciences and other disciplines that depend on observation for their data, the deductive and inductive inferences with which formal logic is concerned are those for which validity depends not on any features of their subject matter but on their form and structure. Logic thus provides criteria by which to recognize valid deductions, correct inductions, and assorted fallacious arguments. As logician W. C. Salmon notes, "there are no precise logical characteristics that delineate incorrect deductions and incorrect inductions; basically it is a psychological matter" (1984, p. 18; emphasis added). Logic is not to be confused with the empirical study of the processes of reasoning, which belongs to psychology (e.g., Braine & Rumain, 1983; Rips, 1994). It also must be distinguished from the art of correct reasoning, which is the practical skill of applying logical principles to a concrete issue or to a particular range of subject matter. Even more sharply, it must be distinguished from the art of persuasion, in which invalid arguments are sometimes more effective than valid ones (Hughes, 1992).

Scholars familiar with the literature on U.S. social policies know that the authors of The Bell Curve did not need the scientific literature on individual and group differences in intelligence and achievement to propose the public policies they propose in that book. Even the authors of The Bell Curve do not unequivocally suggest that their public policy proposals arise necessarily from the scientific data. Scarr (1994-1995) points out that C. Murray had proposed--sans scientific data--essentially the same policies many years ago to a skeptical Congress. The view that antipoverty programs are ineffective, indeed counterproductive, is not a new theme for C. Murray, as Goldberger and Manski (1995), too, remind us. Moreover, as the last two authors note, it is ironic that in his earlier book, Losing Ground (1984), C. Murray's critique emphasized the rationality, or reasoning ability, of the poor, unwed parents, school dropouts, and criminals. Thus he wrote:

> Specifically, I will suggest that changes in incentives that occurred between 1960 and 1970 may be used to explain many of the trends we have been discussing. It is not necessary to invoke the Zeitgeist of the 1960s, or changes in the work ethic, or racial differences, or the complexities of postindustrial economies, in order to explain increasing unemployment among the young, increased dropout from the labor force, or higher rates of illegitimacy and welfare dependency. All were results that could have been predicted (indeed, in some instances were predicted) from the changes that social policy made in

> the rewards and penalties, carrots and sticks, that govern human behavior. All were rational responses to changes in the rules of the game of surviving and getting ahead. . . .
>
> I begin with the proposition that all, poor and not-poor alike, use the same general calculus in arriving at decisions; only the exigencies are different. (C. Murray, 1984, pp. 154-155)

In contrast, Part 3 of The Bell Curve concludes as follows:

> The lesson of this chapter is that large proportions of the people who exhibit the behaviors and problems that dominate the nation's social policy agenda have limited cognitive ability. Often they are near the definition for mental retardation. . . . When the nation seeks to lower unemployment or lower the crime rate or induce welfare mothers to get jobs, the solutions must be judged by their effectiveness with the people most likely to exhibit the problem: the least intelligent people. And with that, we reach the practical questions of policy that will occupy us for the rest of the book. (Herrnstein & C. Murray, 1994, p. 386)

This change in the rationale used to support essentially the same policy directions is a fitting illustration of the proposition that it is values, attitudes, and beliefs--and not rigorous rules of logic--that typically govern the process of drawing policy implications from scientific data. The study of values, attitudes, and beliefs has traditionally been the province of psychology. It is reasonable to conclude, therefore, that psychology--more so than any other single discipline--holds the key to a better understanding of the connections (or lack thereof) between scientific research and public policy.

References

Achen, C. H. (1986). The statistical analysis of quasi-experiments. Berkeley: University of California Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Bjork, R. A. (1994). Institutional impediments to effective training. In D. Druckman & R. A. Bjork (Eds.), Learning, remembering, believing: Enhancing individual and team performance (pp. 295-306). Washington, DC: National Academy Press.

Braine, M. D. S., & Rumain, B. (1983). Logical reasoning. In J. H. Flavell & E. M. Markman (Eds.), Handbook of child psychology: Vol. 3. Cognitive development (4th ed., pp. 263-340). New York: Wiley.

Campbell, D. T., & Boruch, R. F. (1975). Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), Evaluation and experiment: Some critical issues in assessing social programs (pp. 195-285). New York: Academic.

Campbell, D. T., & Erlebacher, A. (1975). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In E. L. Struening & M. Guttentag (Eds.), Handbook of evaluation research (vol. 1, pp.

597-617). Beverly Hills, CA: Sage. Reprinted from J. Hellmuth (Ed.) (1970), Compensatory education: A national debate, Vol. 3: Disadvantaged child. New York: Brunner/Mazel.

Campbell, D. T., & Frey, P. W. (1970). The implications of learning theory for the fade-out of gains from compensatory education. In J. Hellmuth (Ed.), Compensatory education: A national debate, Vol. 3: Disadvantaged child (pp. 455-463). New York: Brunner/Mazel.

Cardon, L. R., Smith, S. D., Fulker, D. W., Kimberling, W. J., Pennington, B. F., & DeFries, J. C. (1994). Quantitative trait locus for reading disability on chromosome 6. Science, 266, 276-279.

Carroll, J. B. (1982). The measurement of intelligence. In R. J. Sternberg (Ed.), Handbook of human intelligence (pp. 29-120). Cambridge, England: Cambridge University Press.

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. Cambridge, England: Cambridge University Press.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 201-219). New York: Macmillan.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton-Mifflin.

Detterman, D. K. (1985). Review of Wechsler Intelligence Scale for Children--Revised. In J. V. Mitchell, Jr., (Ed.), The Ninth Mental Measurements Yearbook (pp.

1715-1716). Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska/University of Nebraska Press.

Dunn, L. C. (1992). Mendel, Gregor. The New Encyclopaedia Britannica: Micropaedia (15th ed., vol. 8, pp. 4-5). Chicago: Encyclopaedia Britannica.

Goldberger, A. S., & Manski, C. F. (1995). Review article: The Bell Curve by Herrnstein and Murray. Journal of Economic Literature, 33, 762-776.

Hamer, D. H., Hu, S., Magnuson, V. L., Hu, N., & Pattatucci, A. M. L. (1993). A linkage between DNA markers on the X chromosome and male sexual orientation. Science, 261, 321-327.

Hart, B., & Risley, T. R. (1995). Meaningful differences in the everyday experience of young American children. Baltimore, MD: Brookes. (Advance copy summary and brochure. Baltimore, MD: Brookes.)

Herrnstein, R. J., & Murray, C. (1994). The bell curve: Intelligence and class structure in American life. New York: Free Press.

Hughes, G. E. (1992). Logic systems. The New Encyclopaedia Britannica: Macropaedia (15th ed., vol. 23, pp. 242-264). Chicago: Encyclopaedia Britannica.

Hume, D. (1739-1749). A treatise of human nature (Book 3, Part 1, Sec. 1). Cited in Salmon, 1984.

Hunt, J. McV. (1961). Intelligence and experience. New York: Ronald.

Jensen, A. R. (1972). Genetics and education. New York: Harper & Row.

Kaufman, A. S. (1994). Intelligence testing with the WISC-III. New York: Wiley.

Knoppers, B. M., & Chadwick, R. (1994). The human genome project: Under an international ethical microscope. Science, 265, 2035-2036.

Lander, E. S., & Schork, N. J. (1994). Genetic dissection of complex traits. Science, 265, 2037-2048.

Laosa, L. M. (1977). Nonbiased assessment of children's abilities: Historical antecedents and current issues. In T. Oakland (Ed.), Psychological and educational assessment of minority children (pp. 1-20). New York: Brunner/Mazel.

Laosa, L. M. (1984). Social policies toward children of diverse ethnic, racial, and language groups in the United States. In H. W. Stevenson & A. E. Siegel (Eds.), Child development research and social policy (pp. 1-109). Chicago: University of Chicago Press.

Laosa, L. M. (1989). Social competence in childhood: Toward a developmental, socioculturally relativistic paradigm. Journal of Applied Developmental Psychology, 10, 447-468. Reprinted from M. W. Kent & J. E. Rolf (Eds.) (1979), Primary prevention of psychopathology: Vol. 3. Social competence in children (pp. 253-279). Hanover, NH: University Press of New England.

Laosa, L. M. (1990). Population generalizability, cultural sensitivity, and ethical dilemmas. In C. B. Fisher & W. W. Tryon (Eds), Ethics in applied developmental psychology: Emerging issues in an emerging field (pp. 227-251). Norwood, NJ: Ablex.

Laosa, L. M. (1991). The cultural context of construct validity and the ethics of generalizability. Early Childhood Research Quarterly, 6, 313-321.

Levy-Lahad, E., et al. (1995). Candidate gene for the chromosome 1 familial Alzheimer's disease locus. Science, 269, 973-977.

Linn, R. L. (1989). Current perspectives and future directions. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 1-10). New York: Macmillan.

Lubin, B., Larsen R. M., & Matarazzo, J. D. (1984). Patterns of psychological test usage in the United States: 1935-1982. American Psychologist, 39, 451-454.

Lubin, B., Larsen R. M., Matarazzo, J. D., & Seever, M. (1985). Psychological test usage patterns in five professional settings. American Psychologist, 40, 857-861.

Mainstream science on intelligence. (1994, December 13). Wall Street Journal, opposite editorial page.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741-749.

Murray, C. (1984). Losing ground: American social policy 1950-1980. New York: Basic Books.

Murray, J. C., et al. (1994). A comprehensive human linkage map with centimorgan density. Science, 265, 2049-2054.

Nelkin, D., & Tancredi, L. (1989). Dangerous diagnostics: The social power of biological information. New York: Basic Books.

Novick, M. (1982). Ability testing: Federal guidelines and professional standards. In A. K. Wigdor & W. R. Garner (Eds.), Ability testing: Uses, consequences, and controversies (Part 2, pp. 70-99). Washington, DC: National Academy Press.

Oakland, T., & Laosa, L. M. (1977). Professional, legislative, and judicial influences on psychoeducational assessment practices in schools. In T. Oakland (Ed.), Psychological and educational assessment of minority children (pp. 21-51). New York: Brunner/Mazel.

Plomin, R. (1995). Molecular genetics and psychology. Current Directions in Psychological Science, 4, 114-117.

Plomin, R., DeFries, J. C., & McClearn, G. E. (1990). Behavioral genetics: A primer (2nd ed.). New York: Freeman.

Plomin, R., & Petrill, S. A. (1995, July). Genetics and intelligence: What's new? Unpublished manuscript.

Ree, M. J., & Earles, J. A. (1991). The stability of g across different methods of estimation. Intelligence, 15, 271-278.

Rips, L. J. (1994). The psychology of proof: Deductive reasoning in human thinking. Cambridge, MA: MIT Press.

Risch, N., & Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science, 268, 1584-1589.

Sattler, J. M. (1982). Assessment of children's intelligence and special abilities (2nd ed.). Boston: Allyn and Bacon.

Sattler, J. M. (1988). Assessment of children (3rd ed.). San Diego, CA: Author.

Salmon, W. C. (1984). Logic (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Scarr, S. (1994-1995, Winter). What is equality? (Review of Herrnstein & C. Murray's The Bell Curve.) Issues in Science and Technology, 82-85.

Snyderman, M., & Rothman, S. (1987). Survey of expert opinion on intelligence and aptitude testing. American Psychologist, 42, 137-144.

Taylor, H. F. (1980). The IQ game: A methodological inquiry into the heredity-environment controversy. New Brunswick, NJ: Rutgers University Press.

Terman, L. M., & Merrill, M. A. (1973). Stanford-Binet Intelligence Scale: Manual for the third revision--Form L-M. Boston: Houghton-Mifflin.

Thorndike, R. L. (Ed.) (1971). Educational measurement (2nd ed.). Washington, DC: American Council on Education.

Wechsler, D. (1958). The measurement and appraisal of adult intelligence (4th ed.). Baltimore, MD: Williams & Wilkins.

Wigdor, A. K., & Garner, W. R. (Eds.) (1982). Ability testing: Uses, consequences, and controversies (Parts 1 and 2). Washington, DC: National Academy Press.

Zigler, E., Abelson, W. D., & Seitz, V. (1973). Motivational factors in the performance of economically disadvantaged children on the Peabody Picture Vocabulary Test. Child Development, 44, 294-303.

Author Note

Portions of this paper were presented by invitation at the 103rd Convention of the American Psychological Association (APA)--in a symposium entitled The Bell Curve Controversy, organized by APA Division 45 (Society for the Psychological Study of Ethnic Minority Issues) and the APA Committee on Ethnic Minority Affairs--New York City, August 11-15, 1995.

Correspondence concerning this paper should be addressed to Luis M. Laosa, Educational Testing Service, Turnbull Hall, 8-R, Rosedale Road, Princeton, NJ 08541.

Footnotes

[1]The WSJ article fails to mention the experts who declined the invitation to sign it.

[2]Personal communication from Donald T. Campbell to Luis M. Laosa, May 18, 1995: letter accompanied by rough-draft document entitled To the Wall Street Journal, which Campbell has kindly permitted me (L.M.L.) to quote.

[3]Heritability coefficients apply only to differences among individuals within groups. Heritability is a statistic intended to ascertain the proportion of phenotypic (observed) differences among individuals in a population that can be attributed to genetic differences among them (e.g., Plomin, DeFries, & McClearn, 1990; see also Taylor, 1980).

[4]A national survey of clinical psychologists (Lubin, Larsen, & Matarazzo, 1984; see also Lubin, Larsen, Matarazzo, & Seever, 1985) found that the Peabody Picture Vocabulary Test (PPVT) is the third most frequently used test of intellectual functioning--the first and second being the Wechsler Intelligence Scale for Children (WISC) and the Wechsler Intelligence Scale for Adults (WAIS). (The Stanford-Binet Intelligence scale, a measure of general intelligence of the same type as the WISC and WAIS, was at one time at least as popular as are now the last two but has lost some of its popularity in recent decades.) As Detterman (1985, p. 1715) observes, users of these tests appear to have developed a general rule: "When a quick measure of IQ will do, use the PPVT, but when a more dependable measure is required, use the WISC [or the Stanford-Binet or WAIS]." After reviewing studies on the PPVT and

intelligence tests, however, Sattler (1982, pp. 271-272; see also Sattler, 1988, pp. 350-351) concluded that, even though PPVT and IQ test scores may be highly correlated, the type of receptive (or recognition) vocabulary ability measured by the PPVT and other picture-vocabulary tests "is related to general intelligence, but it is by no means the same," since such tests measure "only one facet of a child's ability repertoire." Zigler, Abelson, and Seitz (1973) reached a similar conclusion based on other evidence.

[5]It is helpful to distinguish, as Novick (1982) does, the three main participants in the ability-testing process: the test user, using the test for some decision-making purpose; the test producer, who develops and markets or administers and scores the tests; and the test taker, who takes the test by choice, direction, or necessity.

[6]The Standards (AERA, APA, & NCME, 1985) elaborates on its definition as follows: "Differing regression slopes or intercepts are taken to indicate that a test is differentially predictive for the groups at hand. Under these circumstances, a given predictor [e.g., a college admissions test] score yields different criterion [e.g., college grades] predictions for people in different groups and a given criterion . . . yields a different predictor cut score for people in different groups" (p. 13). This is the only approach to defining predictive--or selection--bias adopted in the Standards. The Standards recognizes that "value judgments are always involved in selection decisions, if only implicitly. The question of what value judgments are appropriate in individual applications is not addressed in the Standards" (p. 11).

[7]PKU is a hereditary inability of the body to metabolize normally the amino acid phenylalanine. As a result, the central nervous system is affected--an impairment that symptomatically manifests itself by mental retardation, epileptic seizures, and abnormal brain wave patterns. PKU is transmitted by an autosomal recessive gene. Approximately one in 10,000 newborn infants will show abnormally high plasma phenylalanine levels; of these, about two-thirds will have the classic form of PKU, which, if untreated, will cause severe mental retardation.

[8]DNA, a substance localized in the cells of organisms, constitutes a molecular basis for heredity. DNA markers are bits of DNA that differ among individuals. They are spread throughout the chromosomes and make it possible to identify the location of particular genes on a chromosome (see, for example, Levy-Lahad et al., 1995).

[9]Gregor Mendel (1822-1884), Austrian botanist who laid the mathematical foundation of the science of genetics. His systematic experiments in a small monastery garden led to his discovery of the basic principles of heredity. Mendel crossed varieties of the garden pea that had maintained constant differences in alternative characters such as seed color and seed shape. The monk-scientist theorized that the occurrence of the visible alternative characters of the plants, in the constant varieties and in their descendants, is due to the occurrence of paired elementary units of heredity--now known as genes. He ascertained the statistical consequences of these principles and confirmed them by experiment. His work seems to have had no effect on the biological thinking of his time, although his

publications reached the major libraries in Europe and America; it was rediscovered 40 years later (Dunn, 1992).

[10]It is believed that if multiple genes affect particular behavioral dimensions, or traits, the genetic effect is likely to be continuous, or quantitative. For this reason, genes involved in multigenic systems are called QTL--even if the trait or disorder in question is diagnosed as a dichotomy. For example, although an individual may be diagnosed as having mental retardation or normal intelligence, the genetic effect in the population is continuously distributed, from low to high IQ. Thus, a QTL perspective encourages one to think in terms of quantitative dimensions rather than diagnostic dichotomies (Lander & Schork, 1994; Plomin, 1995).

[11]See, for example, Hamer, Hu, Magnuson, Hu, & Pattatucci, 1993.