1984

2. Struggles and Possibilities: The Use of Tests in Decision Making

Ellis Batten Page
Duke University

2 Struggles and Possibilities: The Use of Tests in Decision Making

Ellis Batten Page
Duke University

What a happy occasion it is to celebrate, as we do in this volume, the establishment of a national Buros Institute of Mental Measurements, located on the campus of the University of Nebraska, in Lincoln. What a culmination of many plans, hopes, and dreams! On such an occasion, we can take a quiet pride in our profession and in the life and accomplishments of one of our colleagues and friends, Oscar Krisen Buros, who with Luella Buros is leaving to us, and our posterity, an institution of integrity to foster the science and practice of testing.

How new all this field really is: According to Stanley and Hopkins (1972, p. 163), the first large-scale testing was done in the City of New York Survey, in 1911. Oscar Buros was 6 years old then, so we can think of most of the astonishing developments in measurement really happening during his lifetime. And the first machine for scoring of answer sheets, the old IBM 805, was developed when Oscar was 30. Many of us can remember, only 20 years ago, many clerical workers reading the dials from these machines and writing the scores as they might be estimated from this analog device. Then these tools also became obsolete as the field was overtaken by optical readers and computer scoring. So Oscar and Luella Buros have witnessed the explosion of testing into a central institution of education, of psychology, of all the social and behavioral sciences. But they have done much more than witness: Their publications have served as a steady center of this growth, and their independence has established a tradition of reputation and honor as a goal, if not always as a realization, of the profession and the practice of testing.

The establishment of such published symposia from the Buros Institute is an important further step. There is a major place for such a forum. I hope these symposia will represent a determined effort to stand apart from the testing giants,

just as Buros did, and to remain independent of federal agencies as well. The Institute, and these symposia, should continue to sponsor solid, sometimes severe criticism of tests and test practices, also as Buros did. They should similarly stand apart from the political huckstering and trend riding, the cheap shots against testing, and apart from the constant distortion of what tests tell us about ourselves and our world.

Of course, the Institute should make full modern use of wordprocessing, automatic mailing, information retrieval, and all the present and future efficiencies of operation becoming available. But hopefully there will remain these steady principles that marked Buros' work, and a similar vision of mental measurement, of how it can help our society to be happier and more productive.

At such a historic time, it is a pleasure to remember the classic words of E. L. Thorndike (1918, p. 16), which serve as a kind of cornerstone for our whole professional and scientific development:

> Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality. Education is concerned with changes in human beings; a change is a difference between two conditions; each of these conditions is known to us only by the products produced by it—things made, words spoken, acts performed, and the like. To measure any of these products means to define its amount in some way so that competent persons will know how large it is, better than they would without measurement. To measure a product well means so to define its amount that competent persons will know how large it is with some precision, and this knowledge will be conveniently recorded and used.

If we have, for our profession, an Apostle's Creed, surely Thorndike has here given it to us. And the last phrase echoes for us: "so that this knowledge will be conveniently recorded and used." *And used.* Aye, there's the rub and the thrust of the testing movement. It is the *use* of testing that has caused its growth from academic curiosity to a billion-dollar industry and that makes it a battle ground today for conflictiong ideologies and the warring of powerful political alliances. In my opinion, technical people in testing cannot go on sidestepping these major battles. Sooner or later, we should recognize publicly what it is that we believe; we should state our beliefs openly for both colleagues and society; and we should counterattack the falsehoods about testing.

Who are these enemies? For one example, let us mention the recent storm of antitesting sentiment surrounding the publication of Gould's (1981) book, *The Mismeasure of Man.* This book follows in the tradition of Leon Kamin's (1973) *The Science and Politics of I.Q.*, the writing of the consumerists Nader and Nairn and of Lewontin, Layzer, and others. Once again, the major media have rushed to approve the new book by Gould and to endorse its claims. A recent New Yorker has an extended piece by one of their science popularizers, in which most test experts are implicitly denigrated and the founders of our discipline are derided and smeared. There are many echoes of these sentiments.

The major media of the Northeastern Seaboard are, of course, considerably more antitesting than is the American mainstream. What of the more conservative press? Although it is part of the conservative tradition to recognize and to accommodate large individual differences, the better-known conservative writers seem daunted by the name-calling and by the technical difficulty of the arguments. Both sides are handicapped by the recondite nature of many of the core proofs of testing. As Garrett Hardin recently commented during a visit at Duke, most opinion leaders and shapers who control our media, of whatever leaning, are highly literate but are "innumerate." Left or right, journalists fail to grasp our technicalities. They believe that our hard-won principles (the best body of theory in the social sciences) are purely a matter of opinion!

Then what about the "numerate" scientists concerned with tests? Those who do speak out often suffer for it and are frustrated again and again by the major media. Consider the experiences of one of our most productive and distinguished defenders of psychometrics, Arthur R. Jensen of Berkeley. Those who know him well can recount some of his harassment and defamation, which, by the way, is still going on. And Richard Herrnstein (1982) of Harvard has written a critique, much of it from his own unhappy treatment, about his efforts to be expressed properly in the major media. His forum is the *Atlantic Monthly,* an intellectual magazine that is highly respected and of general readership but that commands none of the publicity clout of CBS or of the *New York Times* and their multimillion audiences. Some of Herrnstein's (1982) accusation is worth reproducing here:

> Incurably addicted to quantification, I have now searched the daily and the Sunday *New York Times* from 1975 to November 1981 for all book reviews dealing with the IQ. The results speak for themselves. Of the 15 reviews that I found, every one denigrated IQ tests, often vitriolically. All but two of the books reviewed were antitesting, as far as one can tell from the reviews, and were praised for their position. One exception was a book by Arthur Jensen [1980], which happened also to be the only book by a trained psychometrician (psychometrics is the psychological specialty concerned with testing). Jensen's book was panned by a philosopher with no detectable expertise in the subject.
>
> Except for Jensen's book, none of the other major works on testing written by professionals during the period was reviewed. Most remarkably, however, the *Times* published no review by a trained professional. Dozens of literate psychometricians might have commented on the shallowness of the books the *Times* usually chooses to review. But psychometrics is forbidden territory in the *Times*—its books are mostly unreviewed, its discoveries are unreported, and its experts are, from what gets published, unconsulted. Rarely, if ever, in more than a decade, has a specialist published a review of a book on testing in the *Times,* [or in] other national publications that occasionally comment on testing. For no other subject of public concern—not for economic policy, disarmament, welfare reform, nuclear power plants—has the professional outlook on a controversy been so shut off from a voice in the national press. Yet, while public policy on testing may not have the immediacy of a tax cut or a nuclear accident, it ultimately affects everyone [p. 69].

A DOUBLE STANDARD

Herrnstein's (1982) article is a good one, revealing for all concerned with testing and education. Its principal burden is the double standard of treatment of two cases of apparent malfeasance by testing researchers: One of these cases is known widely even to college students; the other is a nonevent, conveniently buried from public awareness. The first, so widely known, concerns the probable falsification of certain twin data by the late, brilliant Sir Cyril Burt. Herrnstein counted at least six stories about this apparent misconduct in the *New York Times* alone. However, as repeatedly noted by scholars of behavior genetics, nothing in Burt's estimates was very deviant from what has been found by other researchers since his reports. Burt's data are, in short, now redundant, and if he did fabricate some of his numbers, he "apparently knew enough to guess correctly" (Herrnstein, 1982, p. 70). But the attacks on him persist, endlessly, and are made central to denigrating not only behavior genetics but our entire field of mental measurement.

The other story will probably be new to many readers and will surely be new to most nonspecialists. In July 1981, Dr. Rick Heber, Director of the Waisman Center of the University of Wisconsin, Madison, and chief adviser to a U.S. president on mental retardation, was convicted in federal court of diverting funds to personal use and was sentenced to 3 years in prison. Heber, it will be remembered, was principal investigator of the much publicized miracle of the environmentalist movement, the "Milwaukee Project." He had proved, he wrote, that it was possible to take 20 children of retarded parents and depressed homes and to raise their true IQs an average of over 30 points, from dull normal to superior in intelligence, by a massive preschool intervention.

What of his results themselves and their claim to scientific seriousness? Eight years before that trial, an article was published for fellow researchers (Page, 1972b), arguing that the Milwaukee Project was, for a number of technical reasons, not scientifically credible. And just before Heber's indictment, another article (Page & Grandon, 1981) carried an intensive criticism of the Project. In brief, we found that the Project, which had never been truly refereed, was extremely shaky, and the explanations of it shifted in ways quite unacceptable in scientific reporting. What evidence was available on follow-up data, moreover, suggested that there was no residual difference between the treatment and control groups on measures, such as school reading tests, which were outside the reach of Project management. The 30 points gain, if it ever existed, had apparently disappeared.

The point here, however, is not to resurrect the Milwaukee Miracle to slay it again but to draw attention to the way that psychometric questions are treated in the media. The earlier "findings" of the Milwaukee Project had been widely noted in the national media. The *Washington Post* believed that it might have "settled once and for all" (sic) the question of heredity versus environment for

the intelligence of slum children. The *New York Times* had reported that the Project "has proved" that IQs could be raised more than 30 points by the methods of Heber and his associates (these quotes cited by Herrnstein, 1982). Wouldn't one suppose, therefore, that the disgrace of the Project leadership deserved some attention? After all, the Milwaukee work had been unique and widely acclaimed in its demonstration of such large environmental effects. And this demonstration had depended on faith in its leadership. Wouldn't the astonishing misconduct of the leadership, then, cast some shadow across such findings, which no one else had obtained?

Not at all. Not a word about the Heber scandal has appeared in the *Times,* the newsweeklies, *Science* magazine, or on national TV. To quote Herrnstein (1982) again,

> The media seem unwilling to publish anything that might challenge the certitude with which editors, politicians, judges, and others insist that we know how to increase measurable intelligence, or that test data "prove," to use *The New York Times*'s word, that a poor environment causes familial retardation [p. 710].

What is the cause of this remarkable double standard? Clearly, it is the ideology of the major media, warmly supportive even of falsehoods favorable to environmentalism, generally condemnatory of individual differences and hence of psychometrics, our field, which persistently and embarrassingly reiterates important and substantial differences in humankind.

Yes, we have our critics, and they have an extraordinary double standard; and they are in very strong positions, affecting the beliefs of everyone: of editors, educators, judges, legislators, federal officers, and the other countless millions who read the national press or listen to the national TV. If we believe in our discipline and its contributions to society, then we had better stand up for ourselves and our field. What, then, do we believe?

THE VALUE OF TESTING

Scientific Value. In our own quiet way, and in our own private literature, there is a strong consensus among us concerning the persisting values of our science and our profession. In an excellent summary of this question, the scientific basis of testing was powerfully defended by Carroll and Horn (1981). They showed our growth to be following the earlier development of physics, in our gathering understanding of intelligence and our strengthening theory.

Poor Alternatives to Testing. Many of our negative reactions to our critics and would-be reformers are similarly shared among ourselves. That the interference of the courts is often ignorant, confused, and damaging is noted by even the mildest of scientific commenters (Bersoff, 1981). And the reforms forced on

testing by outside criticism have, we are largely agreed, been frequently ''non-solutions'' (Reschly, 1981). Such ''unproductive changes'' include the banning of intelligence tests (such as in California) and the use of ''pluralistic norms'' (such as SOMPA; cf. Mercer, 1977). Often aggressive counterattacks to our critics are slipped quietly into our thoughtful articles written for each other. Such a counterattack is well illustrated by the comment of two of our respected colleagues (Carroll & Horn, 1981): ''Indeed, it seems clear to the present authors that far from being abused by overuse, the science of human abilities is underexploited in diagnosis, counseling, and evaluation [p. 1019].''

Fairness to Minorities. For a very important topic, the claim of racial unfairness, the view of experts was well summarized by Cole (1981), when she wrote - that ''we have learned that there is not large-scale, consistent bias against minority groups in the technical validity sense in the major, widely used and widely studied tests [p. 1075].'' This position has been strongly supported by a blue-ribbon panel on testing of the National Academy of Sciences. And a similar conclusion is widely understood for the question of bias in college admissions (Linn, 1982). Indeed, much of the claimed evidence against test validity, for example in employment, has apparently been misunderstood and improperly summarized (especially see Schmidt & Hunter, 1981).

IDEOLOGICAL AND SCIENTIFIC ISSUES

Through many arguments about test practice, however, run deeper currents of contemporary ideology, philosophical, political, and economic. Those who claim an exclusively societal or economic determinism are especially resentful of testing and psychometric research, and what these disciplines show us about the sources of human abilities and personality. In a candid account of the contemporary scene, then, we must not avoid the issue of what science and scientists say about family influences on these traits, both genetic and environmental.

Heritability of Intelligence

Surely we can now say that there is a scientific consensus for the heritability of intelligence, and we can reject the name-calling of those who would say that hereditary influence is a delusion or a hoax. If there is any scholar who honestly questions it, and sincerely seeks evidence, there is a direct solution: Such a person should read—or even just browse—in Fuller and Thompson's (1978) weighty volume, *Foundations of Behavior Genetics.* Absorb the stately march there from fundamental genetic principles to physiology, to neurobiology, to quantitative methods, to the genetics of cognitive and intellectual abilities, to personality and temperament, to mental illness. Loiter, for a while, in the 40

pages of bibliography with their 1500 references. And for those with quantitative curiosity, there are excellent works available (Falconer, 1960; Thompson & Thoday, 1979).

Or, if a scholar seeks further knowledge of the genetic evidence specific to mental measurements, give such a scholar Jensen's (1980) monumental book *Bias in Mental Testing*. Someday this may be more widely recognized as one of the best works ever written on testing, for the serious student of psychometrics. (For other informed appraisals of such evidences, see Bereiter, 1970; Cancro, 1971; Hébert, 1977. And for a nontechnical treatment of the issues, see Jensen, 1981.) But then, how should we convince the lay world outside of the large consensus on this matter of heritability? In 1972, more than 50 scholars from fields bearing on this question published a "Resolution on Scientific Freedom and Heredity," signing the emphatic statement that "*we believe such influences are very strong.*" (Page, 1972c). Of the 50 signers, 60% were in *Who's Who in America,* and four were Nobel laureates. And their statement was published in the most prominent professional journal in psychology. But that testament, too, became a nonevent for the major media to ignore. The national press took no notice of this, nor did CBS when its special, "The IQ Myth," led by Dan Rather, managed, through distortion and omission, to make test scores seem a pure artifact of favored environments. One of the most common responses of informed psychologists and measurement experts is to avoid these questions or, if pressed, to state that these questions are not important for our major concern: the use of tests in decision making. On the contrary, I hope to persuade that such evasions, of such overpoweringly central questions, must lead to waste, futility, and dishonor in our testing field. Indeed, to some extent this has already happened.

Nonetheless, it is curious how blind the media are to this consensus among scientists about the heritability of intelligence. Even Gould's (1981) book, with its strong ideological loading, does not exactly dispute the existence of heritability, though taking exception to nearly every estimate of it. The device used by Gould, and by others before him, is to challenge the *precision* of such an estimate, as if some softness of numbers invalidated the whole pursuit. If a test score is not precise, they seem to affirm, it is useless. If a heritability estimate is not certain, then it is meaningless. One can only imagine the stultifying influence such perfectionism would have had on the growth of any of our sciences? But the clear fact, revealed even in the most polemical criticism to the careful reader, is that there is consensus about the large heritability of general intelligence.

Heritability of Special Abilities

Even among able psychometricians, however, there is much uncertainty about the heritability of specific abilities or achievement measures. To explore this question, a friend and I (Page & Jarjoura, 1979) obtained an unprecedented data

set from one of the two major college testing programs, the respected American College Testing Program (ACT). Our results are briefly outlined here, as bearing on this important and neglected problem affecting many of the tests the schools have so widely adopted. If these measures, too, are loaded with heritability, we should take this fact carefully into consideration.

As is well known, the ACT has four achievement tests, in the four fields of English, Math, Social Studies, and Natural Sciences. From two different years of testing, 1976 and 1978, ACT gathered for us 6800 pairs of twins from the nearly 2 million students who used this excellent program to apply to colleges in those years. These twins were identified from the concordance of surname, birth date, and place of residence (or home phone). Even without knowing which pairs are fraternal or identical, it is possible to do some genetic analysis of such a wonderfully large data base, as long as we are willing to make certain assumptions about same-sexed and opposite-sexed pairs (Scarr-Salapatek, 1971). Here there is little space for technical detail, but let us consider certain findings, displayed in Table 2.1.

Table 2.1 shows results from a factor analysis of the genetic components estimated from our methods (Page & Jarjoura, 1979, p. 115). First, we observe the sizable loadings of the four tests on the principal genetic factor. The heritability estimates of these four tests were all high, by the way, ranging from .64 to .84. That is, each of the four ACT achievement measures showed a substantial heritability in itself. The further question we raised, however, was the extent to which the measures were genetically unique and the extent to which they shared their genetic loadings with the others.

In Table 2.1, Part A shows these loadings of the four measures on the first, unrotated principal factor from the genetic correlations we generated. In Part B of the table, we observe the amount of each of the genetic correlations, which is explained by the principal component. And in Part C we see that there is also a genetic loading specific to each of the four tests (these loadings are in the major diagonal). What is thought provoking, and not often recognized among psychometricians, is that so much of the intercorrelation among such ability and achievement measures should have a unitary factor as its biological source. And it appears that *G* (genetic loading) and *g* (the always observed correlation among diverse mental measures) do indeed have much to do with each other. (See also kinship studies in Behrman, Hrubec, Taubman, & Wales, 1980; Loehlin, Lindzey, & Spuhler, 1975; Loehlin & Nichols, 1976; Martin, 1975.)

From this example, we can score some points against frequent criticisms. One of the repeated claims is that Burt's apparent defection destroyed the basis for any belief in heritability. But obviously, Burt's few disputed twin pairs played no role in this large analysis (nor in numerous other analyses in the United States or abroad). Another strawman from our critics is that we regard intelligence as a "single thing." This claim is clearly false. Here one sees that, even genetically, there are other influences distinct to each trait. Even so, however, here as in all

TABLE 2.1
Principal Factor Analysis of the Genetic Components of Twin Data[a]

Trait	*English*	*Math*	*Soc. St.*	*Nat. Sci.*
	(A) Loadings on principal genetic factor			
	.71	.65	.83	.84
	(B) Component "explained" by principal factor			
English	.50	.46	.59	.59
Math		.43	.54	.55
Soc. St.			.69	.70
Nat. Sci.				.71
	(C) Residual component			
English	.14	.02	−.02	.00
Math		.21	.00	−.02
Soc. St.			.15	.01
Nat. Sci.				.01

[a]From Page & Jarjoura, 1979, p. 115.
[b]Of the total genetic matrix, 81.5% of the variance was explained by the single factor.

matrices of mental measures we see the ubiquitous positive component underlying the whole matrix, which in this analysis is genetic. "Single thing" it is not; indeed, by all estimates, it is based on many gene loci. And psychologically there are surely various subabilities that contribute to it. Still, whatever its nature, *g* does appear, to a greater or lesser extent, in virtually all mental tests.

Still another charge hurled at testers, but denied by our analysis, is that we believe that "genetics is all." Our Table 2.1 clearly rejects any such conclusion, as does the research of everyone else known to us. Indeed, it is the power of behavior genetics that it can best expose those influences that are, indeed, environmental. For example, we may consider the simple declaration that variance of a test is the sum of the genetic variance, the environmental variance, and error:

$$\text{Var(test)} = \text{Var(G)} + \text{Var(E)} + \text{error}. \quad (1)$$

Then it is possible to regard a test score in the way suggested by Fig. 2.1.

For students of testing, this figure seems a most familiar one. From any test, we might infer that the shaded curve represents the variance expectable from error around some true score X'. But let us alter the meaning: Let X' now represent the *genotype,* and the shaded figure represent the variation expected in the *phenotype,* through the operation of a combination of environment and errors of measurement. What such a perspective makes us realize is that, in each one of our mental test scores, we are indeed looking at a genotype, plus other influences. That is, we may consider the individual score to consist of genotype "true" score, the environmental variations around such a genotype, and of course a residual error variance. Indeed, given the enormous amount of research on these matters, we may assert that, for the individual student, most of the

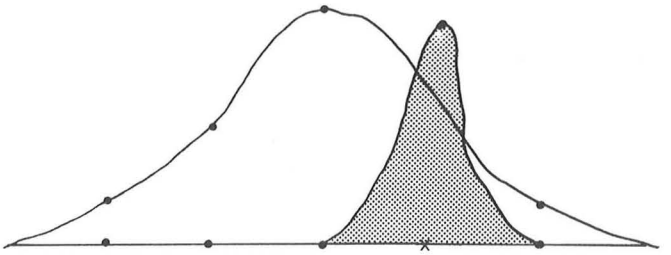

FIG. 2.1. Familiar figure in measurement, applied to either reliability or heritability. *X* may represent the "true score of a test" and the shaded curve represent measurement error. Or it may represent the genotype of a test for an individual and the shaded portion may represent the combination of environmental influences and measurement error.

distance from the mean is difference in genotype, and this is true, whether or not it conforms with the sentiments of CBS or of the *New York Times*. But this assertion in no way denies our pursuit of these environmental causes of test performance. Rather, it clarifies our goal and gives us some methods for identifying the environmental variance without the usual distortion and confounding with an unconsidered background genetic variation.

The formula in Equation (1) is of course very general. A more detailed formula would be the following:

$$\text{Var(IQ)} = \text{Var(E)} + \text{Var(G)} + \text{Cov(G, E)} + \text{Var(G} \times \text{E)} + \text{ error}, \tag{2}$$

where the two new terms represent the covariance and interaction of the genetic and environmental influences. These are surely plausible enough additions to such studies. It is logical that, given the sorting out of social classes, in part caused by differences in ability of the parents, there could be a correlation of genes and environment. And it is also logical that, to some extent, what favors one genotype might not favor another to the same degree.

But such components are difficult to distinguish in twin data and, therefore, are usually neglected in studies of heritability because of mathematical confounding. Yet critics have sometimes used this confounding to disparage any attempts at heritability analysis. One of the critics is an astronomer, who contemptuously referred to the usual methods of human genetics as "numerology," but then himself committed two astonishing logical errors in his mathematical proof (Layzer, 1974). Each of his objections to heritability leads to *reductio ad absurdum*. His policy argument is that all heritability analysis should be curtailed and that we as a society should emphasize only environmental efforts. His prime example of such remediation was the Milwaukee Project (this was, of course, before those investigations were closed and the leaders sent to Federal prison). There were the following two dilemmas: First, GE covariance either exists or it

does not. If it does not exist, then heritability analysis may proceed without it. If it does exist, then Layzer is already granting the argument of Herrnstein (1973) and of others that the upper social classes are already partially sorted for genetic ability in intelligence. Either way, Layzer's practical conclusions are spoiled. His argument about G × E interaction suffers the same fate. If such interaction does not exist, then heritability studies may proceed without it. If it does exist, then, by the very definition of interaction, any marginal improvement in social environment will be, to the extent of that interaction, as unfavorable as it is favorable. (For a more complete treatment of this question, see Page, 1975; and for general treatment of interaction effects in the context of intelligence, see Eaves, Last, Martin, & Jinks, 1977.)

Again, why are such matters worth speaking about, in a volume devoted to tests and decision making? Is it not enough that most able testers acknowledge the truth of heritability and of innate individual differences? Isn't this fact, indeed, something of an embarrassment to testing? Shouldn't we continue, by our passive, noncommittal reaction to these controversies, to paper it over? Isn't it, in fact, almost *bad manners* to raise the question? So it has often been treated, and there is usually, as Herrnstein (1973) points out, a personal and professional cost in resisting the tide of opinion as shaped by the major media.

But these questions are important exactly because our failure to resist such untruths is damaging the reputation of testing and seriously undermining its utility in making decisions. The truth or falsity of our assumptions is crucial to making long-range decisions, by the very nature of scientific decision making. To support this assertion, we turn to the nature of decision making and to the kinds of information required to make an intelligent choice.

DECISION MAKING

We should recognize that a science of decision making has become itself a vast and well-developed field of applied mathematics and statistics with many branches: linear programming, dynamic programming, transportation algorithms, queueing theory, and many other techniques with large implications for behavior science. For a survey of the general field, the reader may see many general texts in operations research (Churchman, Ackoff, & Arnoff, 1957; Hillier & Lieberman, 1974; Trueman, 1974; Wagner, 1969) and increasingly in statistics (Hamburg, 1970; Winkler & Hays, 1975). Some of these methods have been studied for psychology or education (Anderson, 1970; Banghart, 1969; Johnstone, 1974; Kaufman, 1972; Levin, 1975; McNamara, 1971; Novick & Jackson, 1974; Page, 1976, 1978; Page & Canfield, 1975; Page, Jarjoura, & Konopka, 1976; Tillett, 1975; VanDusseldorp, Richardson, & Foley, 1971). A few have brought such methods to bear directly on the use of tests (Cronbach & Gleser, 1965; Edwards, Guttentag, & Snapper, 1975; Page, 1980). In general,

however, there has been little recognition of its importance to educational psychology and its kindred disciplines, and few investigators have applied it to our most serious problems of educational choice.

Decision Analysis. For easy understanding, the science of decision making is often expressed in the notation of *decision analysis,* and the notation is that of an upside-down tree, as shown in Fig. 2.2. The best-known writer in this field is undoubtedly Howard Raiffa (1968), whose approach can be appreciated without extensive mathematics, and can be applied directly in practical situations.

In Fig. 2.2, let us suppose that there is a career choice at stake, such as whether to pursue a premedical career or some other. In this drastically simplified representation, as in many more complex ones, there are just four aspects of choice:

1. Decisions to be made (in squares).
2. Probabilities to be estimated (in circles).
3. Values of the outcomes (numbers at dots).
4. Costs of the choices (small tollgates).

Let us assume that the values of the outcomes are estimated in the same units as the costs at the tollgate. Then such a tree may be automatically solved by applying recursively two rules, beginning at the bottom of the tree and working up:

1. *Probability nodes are averaged,* by multiplying each value by its associated probability, summing across the branches, and carrying the weighted mean up to the node.
2. *Decision nodes are maximized,* by selecting that branch that carries the highest net value. (Costs are subtracted from the value of the relevant branch.)

Tests for Individual Decisions. Decision analysis, then, like most methods used in operations research, suggests our optimal choice, under the assumption of the correctness of our data.

But where do we obtain the number themselves? They are based on some sort of data, either objective or subjective. And the role of tests in forecasting should be closely tied to the probabilities shown. The probabilities of various outcomes, once a decision is taken, must depend on all appropriate information about these outcomes: the experiences of others and the chooser's own abilities, past achievements, economic needs, and the like. For example, suppose that the choice of Plan B is for premedical training, where the payoff ($100,000 a year?) is high but where the general probability of success is only 1 in 5. In the individual case, this probability should be adjusted to the person concerned. Once again, test scores should play an important role in such adjustment, consid-

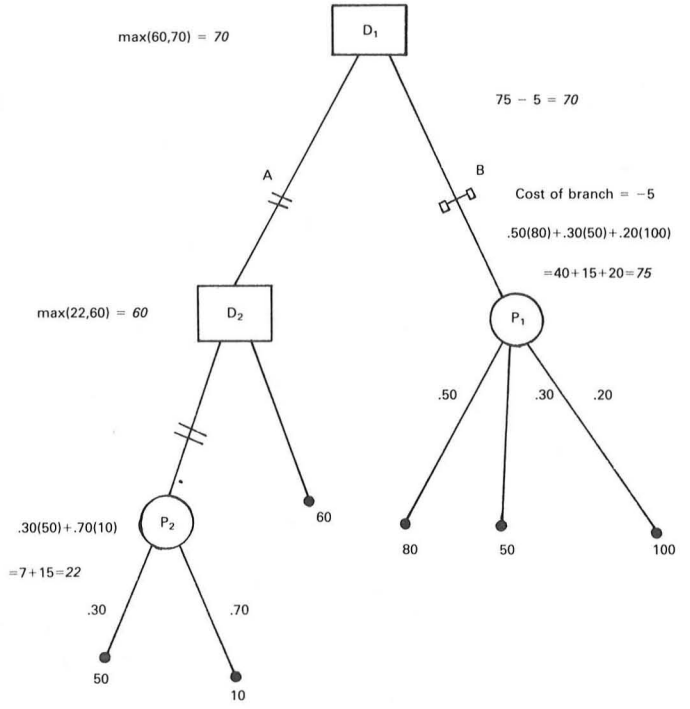

FIG. 2.2. A decision tree. A decision is reached by tracing out the branches as far as possible, assigning values to each terminal node, and probabilities to each branch from a P node. P nodes are then solved (working from the bottom up), by averaging out the branches. And D nodes are solved by folding back all but the most valuable branch as evaluated below each D. For vocations, the probability values are determined by knowledge of both the world and self, as are also the terminal values. Technical procedures can be applied to aid all such determinations. (*Source:* Page, 1974b, p. 71. Reprinted with permission.)

ered together with the background information about others who have gone before.

Consider, then, what great damage is done to decision making, if tests are discredited and not used or if they are eliminated from the tools of decision makers by court order or administrative uncertainty. It is not only the testers who have much at stake in such mistaken elimination of tests; the biggest losers are the students and those who would guide and select them.

From even such a simple model, an immediate realization is that such decision trees become complex, requiring computer assistance in their solution—just as life decisions are indeed complex, yet made quite haphazardly today, without mathematical help. We still await truly competent, computerized advisory systems for such choices, though we have been aware of the need for some years

(Page, 1974b), and some working research models were established in the past (Katz, 1966).

Tests for Program Decisions. Let us look at Fig. 2.2 from another viewpoint, as though we were administrators and the decision were between two programs, here labeled A and B. Suppose Program B seems to produce higher average values, where these are measured in terms of test scores, but our data are from a national study, where there is confounding of tests with school practices and with the SES variables of the communities. What we face, again, is that decision sciences must depend not on naive correlational data but on *production functions* of the treatment variables. If this seems an obvious point, then it has been seriously neglected in the social planning of the past several decades, and its neglect has led us to one disillusionment after another in the world of educational research and development (cf. Page, 1972a).

Scores as Production Functions. In our desire to use tests in planning, we are often blocked when we must choose among educational programs. Choosing a criterion test then becomes troubled. Suppose one program relies more on a textbook and the other more on films. Then it will be very difficult to construct a test that will not be biased toward one outcome or the other. Quite understandably, in such a situation, we often wisely choose tests that are not so close to the programs. We may, rather, choose a selection of standardized tests of global ability or achievement: in English, for example, or in math, social studies, or natural sciences. But wait, these are the very tests we found to be heavily loaded on the same *g* factor (general ability). Even more disturbingly, they are loaded on the same *G* factor (general *genetic* ability). And when we employ pre–post testing with such measures, the change scores have well-known problems. Are we really expected to detect the effects of programs through such measures of general (and even genetic) ability?

Yes, in general we must, for there seem to be few defensible alternatives. We have mentioned the experimental bias of tests designed explicitly for the comparisons, and these (even where available) have many problems beyond such built-in program biases. Tests that are called "criterion referenced" frequently exhibit these problems. We have long seen much literature for and against such criterion referencing, and some excellent consideration and debate have occurred (for example, by Julian Stanley, Robert Ebel, Roger Lennon, and Frank Womer). For an extended period Dr. Womer directed the massive National Assessment of Educational Progress, which was dedicated, at least originally, to the criterion-referencing philosophy (also see Page, 1982, on this philosophy). For research questions about programs, such issues have a special bite.

Special Versus General Tests of Achievement. Let us briefly summarize our dilemma: On one hand, it is fairly easy to write tests that measure some very

limited body of knowledge (e.g., the new vocabulary taught in a specific lesson). Here indeed we can show marked change from before an instruction to after. On the other hand, a small handful of words will have practically no visible effect on one's ability to read general matter—and this is the goal we really cherish for major decisions. If we test only the explicit program content, we may be acting out something like the "drunkard's search," which the philosopher Abraham Kaplan used to tell us about at UCLA. The drunkard was feeling around under a lamppost and was asked what he was looking for. "I dropped my key." Where did you drop it? "Over there." But if you dropped it over there, why are you looking for it over here, under the lamppost? "This is where the light is."

We can, after all, develop a test for the lesson just past, which may show us how we improved. That is where the light is. But the most important outcomes of education often seem like the lost key, beyond our reach, over there in the dark.

Is there a way out of this problem? Yes, if we have sufficient numbers and sufficient random assignment and accurate enough predictive control variables, then our standard errors of the means will be small enough to permit comparisons that are meaningful for such standard testing programs. Such conditions, however, hold in probably less than 1% of the evaluation situations that face the psychometric researcher.

Showing Environmental Effect. The problem is not hopeless. If we have, indeed, important variables, sufficient cases, and solid models, we may be able to show these important environmental influences in a helpful light. Let us consider two findings from recent research on the applied issue of private and public schooling.

Our first case illustrates the danger of failing to provide for large individual differences (in g or in G). Coleman, Hoffer, and Kilgore (1981) had claimed very prominently that, even after "controlling" for effects of family, they found a striking superiority of the private schools in the United States in the educational achievement of the huge sample from High School and Beyond. In a reanalysis of the data, however, this time including six brief subtests of mental ability (mostly nonverbal and relatively school-free), we found that any residual effect of private school was less than 0.5% of the variance in student achievement (Page, 1981; Page & Keith, 1981). Thus a claimed environmental effect largely disappeared when student input was weighed into the test. This is, of course, a common enough result when such variables are included—which has apparently led some to wish to avoid measuring intelligence in such research.

The second findings, from the same debate, had a more optimistic outcome, as shown in Fig. 2.3. In Fig. 2.3, we observe some major student variables, such as family background, race, and general ability, which are understandably loaded with parental influences, both genetic and environmental, and largely beyond the control of the school system. But here we also introduced the amount of homework the student did, as a causal variable for the general achievement of

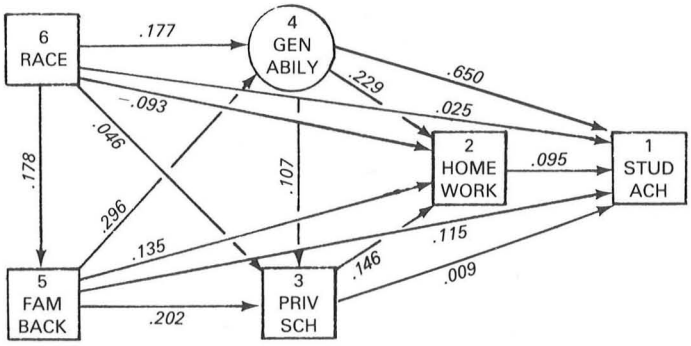

FIG. 2.3. Explaining student achievement of students in private and public high schools. After allowing for background variables, homework still explains 3% of test achievement. (*Source:* Page & Keith, 1981. Reprinted with permission.)

reading and math. Clearly, from the paths shown, our major background variables were fundamental in explaining test achievement; and the special control of "general ability" (a factor score made from two short vocabulary tests and four nonverbal tests) was the most influential of all. Yet the homework does shine through, explaining 3% of the variance in achievement even after controlling for background variables. The eternal verities of educational psychology still stand: After ability, time spent on task does make the most difference, and our standard tests, even loaded as they are with heritability, *can* show that such time matters. Indeed, in this case of school comparison, homework also helps explain about half of the tiny effect of private schools.

Tim Keith and I believe that homework, then, is a major variable that all schools should emphasize, one that could truly improve performance. Keith's (1982) separate article shows this homework effect even more clearly for student *grades:* There is, in fact, a possible compensation for low ability shown in this study of grades, with the low-ability hard worker actually catching up with the high-ability nonworker in such school performance. Keith's remarkable graph is shown in Fig. 2.4.

But another problem of practical decision making is illustrated in this homework question. I have talked about these results with various groups of policy people: school boards, legislators, practicing administrators, equal opportunity officials, teachers, and even governors active in education. The idea of increasing homework seems to have no lobbies! To the contrary, there is often an embarrassed silence (and the facts are indeed embarrassing, with the average senior doing less than 4 hours of homework each week, in all classes combined). Some educators have even denigrated the homework question altogether, speaking of "meaningless drill" and the like. Clearly, far more than our psychometric research enters into educational policy! But this case does illustrate how test

information may improve our knowledge of basic issues, and our understanding, if not always our application, of practical issues.

Heritability and Program Research. Our general neglect of heritability has led to research handicaps that may unfortunately hinder our understanding of some policy issues. In order to guide curricular change, we should know which variables are relatively more influenced by family variables and which more influenced by schools. But our usual research strategies, with no kinship controls, do not often permit this distinction. Given large samples of twins and siblings, however, and item information across achievement tests, we could do heritability analysis on *each item.* Or, if zygosity were not known for the twin pairs, we could analyze which items were more influenced by home or school, and various analyses of these results could in turn illuminate areas for greater curricular attention in those schools showing such deficits.

Still another application of such techniques could be in matters of national assessment, where we seek to track the national performance of student generations and to study the changes from one generation to another. For example, there remain large questions about the causes of the decline of standardized test scores over the past 15 years or so. One real possibility—that declines were caused by shifting ability levels of parents—was never really explored. Yet item analysis of the SAT scores, using the huge available samples of twins, might cast

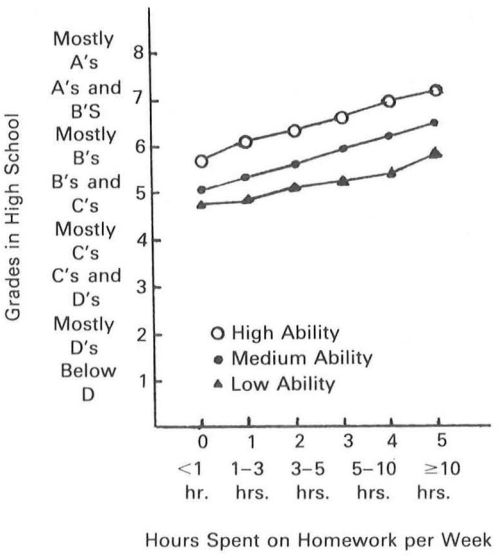

FIG. 2.4. Mean grades in high school as a function of time spent on homework and of ability. (*Source:* Keith, 1982. Reprinted with permission.)

some light on the question, through the following reasoning: Reused items may be measured at two points in time and their gain or loss reported. The twin correlations of such reused items may be also discovered. Then if the more family-influenced items are those in which there is greatest decline, the inference would be that the decline was more likely caused within the home than within the school; and the conclusions would be quite different from those in the contrary case. Conceivably, this exploration would not be very productive (we would soon find out), but it would open a major line of investigation. And it was a thesis that would be very easy and inexpensive to explore. A major cluster of hypotheses remained unstudied. Once again, our psychometric understandings are frustrated by our current political and ideological commitments. And we have failed to make adequate use of the psychometric information available to us in our search for improved social strategies.

Decision Making and Ratio Scales. One apparent problem of test scores for decision making is the following: Most scientific strategies for optimizing decisions require that benefits be measured on some *absolute scale of values.* In many decision techniques (such as certain kinds of dynamic programming), one develops a ratio of costs and benefits for each alternative choice, a ratio that makes no sense unless both costs and benefits have some recognized zero points. Even in simple decision trees like that in Fig. 2.2, where costs are used there must be some way of equating costs and benefits; they must be translated to the same scale. But in mental measurement, we take most of our test scores to be interval scaled, not ratio scaled. How may this difficulty be overcome, so that the most important outcomes of education may be appropriately studied?

This question has been considered elsewhere, but some general answers may be suggested here. Any time we consider *change in scores* then we have, indeed, a ratio scale, for no change will be zero; two points will be twice the value of one point, etc. Now, as we know, change scores have their own problems, because the error variances are additive, whereas the subtraction of one score from the other eliminates from the result most of the variance in the true scores. But if we use *group* change scores, as we often will in program decisions, then indeed the errors of measurement are made very small as the number of observations grows large; and our analysis may proceed.

Often, of course, we will not have repeated measures on the same group but will have some other groups that may be regarded as controls for comparison purposes in our multivariate studies. Here, again, a zero point may be established as the mean made by the relatively "untreated" control, and a production function may be estimated as a relation between possible alternatives and the growth in such means. This should not give the impression that all such questions of zero points are easily resolved but that they can become tractable for many practical purposes in scientific decision making. And we are currently taking little advantage of such strategies.

TEST SCORES AND DEEPER VALUES

Test scores, we have assumed, measure those outcomes for which we most depend on our schools. The scores, then, stand for social values that we highly esteem. Yet strangely little attention has been given to the placing of these test values in some higher framework.

Suppose we ask the simplest curricular question: For example, should we double the time for mathematics in a certain grade, at the expense of some other course of study, such as history? How could we obtain evidence to help guide us in this decision? It is striking that, after 70 years of using test scores and a century of behavioral science, we still have no commonly accepted way of combining such test scores or of trading them off against each other.

The Bentee. A decade ago, some of us studied this question, with the concern of being able to use test scores as production functions (Page, 1972d, 1973, 1974a, 1976, 1980; Page & Breen, 1974a, 1974b). In this work, we felt it necessary to invent a unit of measurement of educational benefit, called the *bentee,* for *benefit T-score.* An illustration of the bentee is shown in Fig. 2.5.

In this figure, we note that the bentee represents the highest educational value, and the branches beneath it stand for seven major branches of educational gain, ranging from the verbal, quantitative, social sciences, and natural sciences through esthetic learning, matters of the body (such as sports, health) to the "personality" (which may include citizenship and moral and spiritual learning where these are deemed appropriate). Each of these major branches may be itself divided into subdivisions. In the present figure, only one, *verbal,* has been divided into seven exhaustive areas. And one of these in turn, *literature,* has been divided. And the tree branches down through *poetic analysis* and *poetic meter,* to *iambic pentameter,* the great verse metric that has been the medium of Shakespeare and of many of our greatest English poets. Recognition of iambic pentameter, then, may be an explicit goal of instruction for good English students; it would be a suitable topic for a test item or for an operational objective in instruction. In these steps, we observe that the tree reaches from the highest philosophical and social values, through only a few steps, to the lowliest and most concrete behavioral objective. Surprisingly, climbing down this tree, the educational philosopher may actually be able to converse (chatter?) with the educational psychologist, who may be occupied with behavior modification techniques.

But how is the actual "evaluation" carried on? Having investigated two methods, we believe that a "token" method may be suitable for most curricular purposes: In this method, appropriate judges, acting individually, apportion 100 tokens (such as poker chips) among the half-dozen divisions at each branch. The method may be applied recursively, at any level of the tree, and by judges chosen as appropriate to that level. At the top, it might be educational leaders or simply

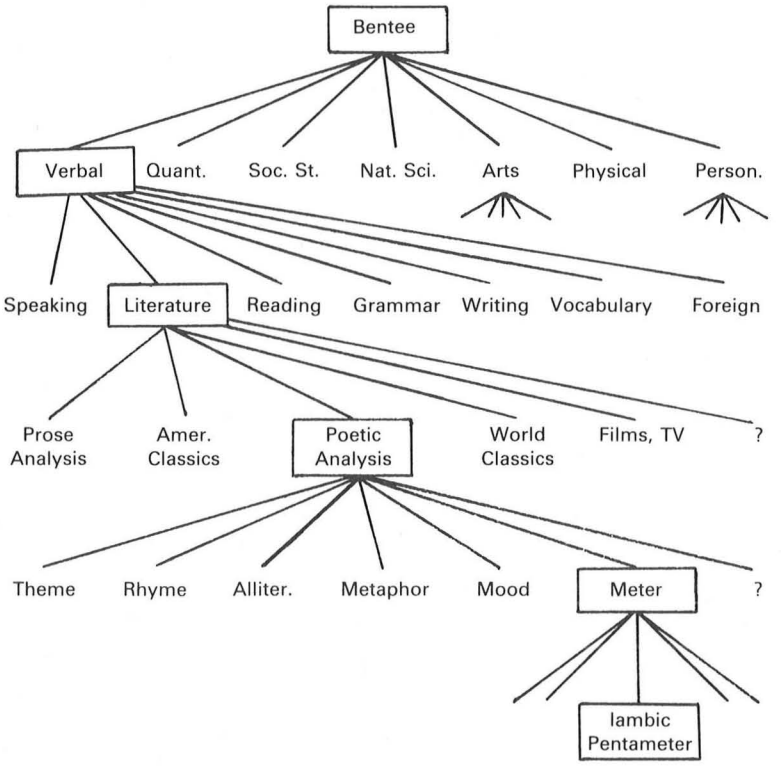

FIG. 2.5. The recursive nature of the bentee method. As analysis moves from the general to the specific, a shift is made from societal to expert opinion and from value space to test space. (*Source:* E. B. Page, 1974. Reprinted with permission.)

informed citizens. At the lower levels, it might be subject matter specialists or future employers (in training situations). These trees may be adapted for any new program of study with its own nodes and branches.

Such a tree has a fairly clear relation to the use of test scores in decision making. Where we have test scores for the various branches (such as English, math, social studies, natural sciences), we may apportion our tokens according to our beliefs in the relative benefits of these accomplishments. And our weightings may vary with the individual concerned (the general student may have a different weighting vector from the premed) or with the program under study. But once such judgments are established, then we may proceed to *evaluate* the educational accomplishment of individuals, of groups, and of programs. By adopting changes in such bentees as our objectives, we may plot our production functions as a relation between decision alternatives and the values that we seek to optimize. Given such methods, we may employ much more frequently the well-developed techniques of the decision sciences in our own studies of policy. (The

reader is directed to related literature: For a technical approach different from the bentee, see Dalkey, 1969. For a deeper understanding of means-end analysis, see Churchman, 1961. For a classic treatment of personnel decisions and test scores, see Cronbach & Gleser, 1965; and for the most advanced general treatment of multiple objectives, see Keeney & Raiffa, 1976.)

PRODUCTION FUNCTIONS AND CAUSAL RESEARCH

We have already noted that "production functions" must involve more than incidental relationships between variables. When we seek to "optimize" some benefits from our decisions, we must depend on the assumption of a *causal relation* between the decision alternatives and the desired benefit. For example, suppose we note, as many researchers have, a recurring agreement between child intelligence and family income. If we believe that this relation is causal, then we naturally predict that when we change family income, we will correspondingly change child intelligence, at least to some limited degree. Programs to eliminate poverty, therefore, according to this reasoning, should have a strong influence on reducing school failure.

Or if we believe that such intelligence is a causal outcome of time spent with the child by a well-intentioned adult, then we will predict that programs such as Head Start will have a clearly beneficial effect on future performance of participating children. Many programs of recent decades have, in fact, been constructed on the assumption that observed correlations of this sort represented strong causal relations. The disappointment about such programs results from the ambiguous and debatable outcomes actually observed. (For a sharp disappointment in a major experiment, see Page, 1972a.) We do not need to resolve these issues themselves to understand the need for some improved methods of policy study. The most important improvement seems to be this: We must routinely seek out data that will permit us to establish *causal models* explaining the maximum amount of variance possible of those variables that we wish to optimize. This means, in the first place, that we indeed have such models and, in the second, that we systematically collect the information that will maximize our knowledge. The first requirement implies that we must turn to *path analysis* to make explicit our causal models. Figure 2.3 shows such a model for exactly such a purpose, here seeking a causal influence of homework time on test achievement. The second requirement implies that we should emphasize the use of comprehensive data sets, rich with the correlates, whether from school, society, or family, that most aid in causal explanation of our outcomes of interest. In Fig. 2.3, then, we truly wish to know the effect of homework time on achievement; but we do not wish to be deceived by the correlates of race, SES, or other school variables in establishing our "production function." But if we did not collect these background characteristics (including intelligence) or if we did not com-

bine them properly into our causal model, we would be utterly deceived about the effect of homework (just as Coleman, as mentioned earlier, was deceived in his claimed effect of private schools).

Path Analysis. As a testing profession, then, interested in policy decisions, we must turn to the rich discipline that is now the center for policy research in most social sciences. This is the field of path analysis, introduced by Sewall Wright (1921) some 6 decades ago. In its wandering route, it has come from genetics, to economics, to sociology, to education and psychology and is now found at the heart of many of the research journals in these fields. The number of textbooks about path analysis has rapidly increased in recent years, and these have improved in complexity and quality (Aigner & Goldberger, 1977; Blalock, 1971; Duncan, 1975; Heise, 1975; Kenny, 1979; Kerlinger & Pedhazur, 1973, ch. 11; Li, 1975; Pedhazur, 1982, chs. 15–16; Taubman, 1977).

There is excellent research discipline in using these models. Because they are explicitly causal, their use forces us to specify our hypotheses about the presence and direction of causal influences and strongly encourages us to employ in our models whatever variables we have available that may illuminate our interests. Drawing and publishing such a model, moreover, forces us to put "up front" our assumptions about these influences. If we have left out measures of intelligence, say, or family influence, then this will be apparent in our model. Or if we have placed variables in the wrong order, thus distorting the influences, this too will be apparent to our readers, whether they are allies or critics. These considerations, clearly, have huge meanings for debates about policy decisions. Indeed, without such considerations of background influences, it would be difficult, if not impossible, to plot out any ratios for costs versus anticipated benefits.

Comprehensive Data Sets. The second major requirement for such causal reasoning is the availability and use of large data sets containing the information necessary for causal inference and estimation. High School and Beyond is probably the most pertinent and available data set for many current concerns. It will be still more valuable as the follow-ups are completed and distributed in 1983 and beyond (current tapes are available through the National Center for Education Statistics, U.S. Department of Education, Washington, D.C.). A splendid data set is also available in the predecessor to HSB, the National Longitudinal Study of 1972, with its four follow-ups (also available from the NCES in Washington). Still another valuable set of tapes may be obtained from the U.S. Department of Labor, dealing more with work and later life and less with the high school years. But each of these data sets lacks something of great importance in family background and many other matters that might be of large interest for many particular policy questions.

Still, such data sets are much more powerful than many realize, even when they appear to lack certain variables of prime concern. Advanced path techniques

involving *unmeasured variables* or *latent variables* can often generate new factors much closer to the variables of real concern. For instance, we generated a relatively school-free "mental ability" from factor analyzing a set of short mental tests (Page & Keith, 1981). Others have similarly constructed factors of "self-concept" from a collection of items about attitude. HSB already supplies an excellent SES scale from a weighted sum of many relevant questions about education, occupation, home, and other factors. In general, then, a rich data set can be much more than the simple sum of its parts.

CONCLUSIONS

From this analysis, there are some strong inferences to draw about the use of tests in decision making, and we briefly summarize them here:

1. Test professionals and test users should stop being placed on the defensive by ill-informed and polemical critics. We should reassert, firmly and publicly, the many virtues of testing and the superiority of making decisions using tests, compared with those made without tests.
2. We should insist that psychometricians and others depending on tests be heard in the major media when tests are discussed.
3. We should stop being apologetic about the reality that tests do, in part, show genetic influences and other family influences as well as social environmental influences. These are in fact part of their purpose.
4. We should cite frequently the research on the alleged biases of the most widely used standardized tests of ability and achievement. In general, the conclusions are similar to those of the blue-ribbon National Academy of Sciences panel: When properly used, *tests are not biased against English-speaking U.S. minority groups*.
5. The measurement of intelligence is one of the greatest achievements in all behavior science. The attempts to eliminate it from consideration at many decision points (such as selection for certain programs, schools, colleges, and professions) are not in the best interests of education nor of society as a whole.
6. When using tests in research on achievement, we should often lean toward the available standardized instruments, especially when these may be treated securely.
7. In such research situations, we should commonly control for the entering ability of the students. It is often fallacious to make program selection without such controls and may lead us to wasteful and disillusioning programs.
8. To assist in making decisions, test researchers should become more familiar with methods from the decision sciences, which permit technical analysis of projected costs and benefits.

9. To make use of such decision models, test researchers should translate scores, where necessary, into useful *values* to serve as production functions.

10. But to use such production functions, we must look closely at the underlying causal relations of the variables (such as achievement) that we wish to optimize. We should design these relations into explicitly causal models in path analysis.

11. If researchers look only at the variables of narrow interest, they will often be deceived by what Simon called "spurious correlation." Rather, the explanatory variables must be expanded to control, as much as possible, for background correlates of both programs and outcomes.

12. Many of such correlates will be found strongly active in family influences. To study such family influences, wherever possible researchers should look to twin pairs and other sibling and kinship relations, together with their degree of kinship (e.g., if known, whether twins are identical or fraternal).

13. It is important that government agencies, large testing corporations, and other collectors of data recognize the explanatory power of such family information and collect such variables into data sets wherever feasible.

14. And it is, finally, important that data sets be made inexpensively available to researchers, so that the causal study of human achievement may proceed in as open and active and public an environment as we can create.

Now it should be evident why this chapter is called *Struggles and Possibilities*. Testing is struggling under attacks by many enemies, operating from many motives and conceptions, often incorrect. And testing is also under constant criticism from its friends. It is friendly criticism, of course, that most characterizes the scientific enterprise and the tradition of Oscar Buros, as editor and model for this Institute that we celebrate in this volume. It is this ferment of friendly and informed criticism that has fostered the splendid growth of our field in its theoretical structure and in the construction and use of tests. These too are struggles, and they are essential to the continuing evolution of testing. Surely, the Buros Institute will continue this tradition of sharp and searching criticism by its most knowledgeable friends.

We must call upon ourselves, as well, to defend our field firmly against the defamations and uninformed assaults by its enemies. If we are faithful to the scientific tradition of open scientific debate and self-criticism, then testing will continue to grow and flourish, just as it has during the Buros' shared lifetime of work. But let us, and the Institute, firmly and courageously *take sides*.

Our field, after all, is probably the soundest structurally of any in the social and behavioral sciences. It is probably the most useful for decision making, for individuals and for social programs. For its past accomplishments, it probably has the smallest amount of apology to make—though it will surely be transformed in each succeeding generation as more is learned. Let us celebrate the field as we celebrate the Buros Institute. Perhaps the Institute might prominently

display on its wall those famous lines from Shakespeare, now as applicable to our discipline as to ourselves as individuals:

> This above all: To thine own self be true,
> And it must follow, as the night the day,
> Thou canst not then be false to any man.

REFERENCES

Aigner, D. J., & Goldberger, A. S. (Eds.). *Latent variables in socioeconomic models.* Amsterdam: North-Holland, 1977.

Anderson, G. E., Jr. Operations research: A missing link. *Educational Researcher,* March 1970, *21,* 1–3.

Banghart, F. *Educational systems analysis.* Toronto: Collier-Macmillan, 1969.

Behrman, J. R., Hrubec, Z., Taubman, P., & Wales, T. J. *Socioeconomic success: A study of the effects of genetic endowments, family environment, and schooling.* Amsterdam: North-Holland, 1980.

Bereiter, C. Genetics and educability: Educational implications of the Jensen debate. In J. Hellmuth (Ed.), *Disadvantaged child* (Vol. 3). New York: Brunner-Mazel, 1970.

Bersoff, D. N. Testing and the law. *American Psychologist,* 1981, *36*(10), 1047–1056.

Blalock, H. M., Jr. (Ed.). *Causal models in the social sciences.* Chicago: Aldine, 1971.

Cancro, R. (Ed.). *Intelligence: Genetic and environmental influences.* New York: Grune & Stratton, 1971.

Carroll, J. B., & Horn, J. L. On the scientific basis of ability testing. *American Psychologist,* 1981, *36*(10), 1012–1020.

Churchman, C. W. *Prediction and optimal decisions: Philosophical issues of a science of values.* Englewood Cliffs, N.J.: Prentice-Hall, 1961.

Churchman, C. W., Ackoff, R. L., & Arnoff, S. L. *Introduction to operations research.* New York: Wiley, 1957.

Cole, N. S. Bias in testing. *American Psychologist,* 1981, *36*(10), 1067–1077.

Coleman, J., Hoffer, T., & Kilgore, S. *Public and private schools.* A report to the National Center for Education Statistics by the National Opinion Research Center. University of Chicago, March 1981.

Cronbach, L. J., & Gleser, G. *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press, 1965.

Dalkey, N. Analyses from a group opinion study. *Futures,* December 1969, *1,* 541–551.

Duncan, O. D. *Introduction to structural equation models.* New York: Academic, 1975.

Eaves, L. J., Last, K., Martin, N. G., & Jinks, J. L. A progressive approach to non-additivity and genotype-environmental covariance in the analysis of human differences. *British Journal of Mathematical and Statistical Psychology,* 1977, *30,* 1–42.

Edwards, W., Guttentag, M., & Snapper, K. A decision-theoretic approach to evaluation research. In E. L. Struening, & M. Guttentag (Eds.), *Handbook of evaluation research* (Vol. 1). Beverly Hills, Calif.: SAGE, 1975. Ch. 8, pp. 139–182.

Falconer, D. S. *Introduction to quantitative genetics.* New York: Ronald Press, 1960.

Fuller, J. L., & Thompson, W. R. *Foundations of behavior genetics.* St. Louis, Mo.: Mosby, 1978.

Gould, S. J. *The mismeasure of man.* New York: Norton, 1981.

Hamburg, M. *Statistical analysis for decision making.* New York: Harcourt, Brace & World, 1970.

Hébert, J. P. *Race et intelligence.* Paris: Copernic, 1977.

Heise, D. R. *Causal analysis.* New York: Wiley, 1975.

Herrnstein, R. J. *IQ in the meritocracy.* Boston: Little, Brown, 1973.

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

Hillier, F. S., & Lieberman, G. J. *Introduction to operations research* (2nd ed.). San Francisco: Holden-Day, 1974.

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

Jensen, A. R. *Straight talk about mental tests.* New York: Macmillan–Free Press, 1981.

Johnstone, J. N. Mathematical models developed for use in educational planning: A review. *Review of Educational Research,* 1974, *44*(2), 177–201.

Kamin, L. *The science and politics of IQ.* New York: Wiley, 1973.

Katz, M. R. A model of guidance for career decision-making. *Vocational Guidance Quarterly,* September 1966, 2–10.

Kaufman, R. *Educational system planning.* Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Keeney, R. L., & Raiffa, H. *Decisions with multiple objectives: Preferences and value tradeoffs.* New York: Wiley, 1976.

Keith, T. Z. Time spent on homework and high school grades: A large-sample path analysis. *Journal of Educational Psychology,* 1982, *74,* 248–253.

Kenny, D. A. *Correlation and causality.* New York: Wiley–Interscience, 1979.

Kerlinger, F. N., & Pedhazur, E. J. *Multiple regression in behavioral research.* New York: Holt, Rinehart & Winston, 1973.

Layzer, D. Heritability analyses of IQ scores: Science or numerology? *Science,* 1974, *183,* 1259–1266.

Levin, H. M. Cost-effectiveness analysis in evaluation research. In M. Guttentag & E. L. Struening (Eds.), *Handbook of evaluation research* (Vol. 2). Beverly Hills, Calif.: SAGE, 1975. Ch. 5, pp. 89–122.

Li, C. C. *Path analysis: A primer.* Pacific Grove, Calif.: Boxwood, 1975.

Linn, R. L. Admissions testing on trial. *American Psychologist,* 1982, *34*(3), 279–291.

Loehlin, J. C., Lindzey, G., & Spuhler, J. N. *Race differences in intelligence.* San Francisco: Freeman, 1975.

Loehlin, J. C., & Nichols, R. C. *Heredity, environment, and personality: A study of 850 sets of twins.* Austin: University of Texas Press, 1976.

Martin, N. G. The inheritance of scholastic abilities in a sample of twins: II. Genetical analysis of examination results. *Annals of Human Genetics, London,* 1975, *39,* 219–229.

McNamara, J. F. Mathematical programming models in educational planning. *Review of Educational Research,* 1971, *41*(5), 419–446.

Mercer, J. R. *Labelling the mentally retarded.* Berkeley: University of California Press, 1977.

Novick, M. R., & Jackson, P. H. *Statistical methods for educational and psychological research.* New York: McGraw-Hill, 1974.

Page, E. B. How we all failed in performance contracting. *Educational Psychologist,* 1972, *9,* 40–42. (a)

Page, E. B. Miracle in Milwaukee: Raising the I.Q. *Educational Researcher,* 1972, *1*(10), 8–15. (b)

Page, E. B. Resolution on scientific freedom and heredity. *American Psychologist,* 1972, *27*(7), 660–661. (c)

Page, E. B. Seeking a measure of general educational advancement: The bentee. *Journal of Educational Measurement,* 1972, *9*(1), 33–43. (d)

Page, E. B. Effects of higher education: Outcomes, values, or benefits. In L. C. Solmon & P. Taubman (Eds.), *Does college matter? Some evidence on the impacts of higher education.* New York: Academic, 1973. Pp. 159–172.

Page, E. B. 'Top-down' trees of educational values. *Educational and Psychological Measurement,* 1974, *34*(3), 573–584. (a)

Page, E. B. Problems and perspectives in measuring vocational maturity. In D. E. Super (Ed.), *Measuring vocational maturity for counseling and evaluation.* Washington, D. C.: Monograph of the National Vocational Guidance Association, 1974. (b)

Page, E. B. Heritability of intelligence: Methodological questions. Technical comment, *Science,* 13 June 1975, *188*(4193), 1126–1128.

Page, E. B. The optimization of educational values in Navy curriculum design. *Proceedings of the American Statistical Association: Social Statistics,* Part II: 1976, 655–659.

Page, E. B. Should educational evaluation be more objective or more subjective? More objective! *Educational Evaluation and Policy Analysis,* 1978, 1(1), 5–6

Page, E. B. Tests and decisions for the handicapped: A guide to evaluation under the new laws. Special issue: A monograph, *Journal of Special Education,* Winter 1980, *14*(4).

Page, E. B. The media, technical analysis, and the data feast: A response to Coleman. *Educational Researcher,* 1981, *10*(7), 21–23.

Page, E. B. *Rethinking the principles of national assessment: Towards a more useful and higher quality knowledge base for education.* Report commissioned by the National Institute of Education. In ERIC, 1982.

Page, E. B., & Breen, T. F., III. Educational values for measurement technology: Some theory and data. In W. E. Coffman (Ed.), *Frontiers in educational measurement and information processing.* Boston: Houghton-Mifflin, 1974. Ch. 3, pp. 13–30. (a)

Page, E. B., & Breen, T. F., III. Factor analysis of educational values across two methods of judgment. *Proceedings of the 15th Interamerican Congress of Psychology* (Bogotá, Colombia), 1974, pp. 106–107. (b)

Page, E. B., & Canfield, J. *Design of Navy course structure through a dynamic programming algorithm.* Report for the U.S. Navy Personnel R. & D. Center, San Diego, Calif., June 1975.

Page, E. B., & Grandon, G. M. Massive intervention and child intelligence: The Milwaukee Project in critical perspective. *Journal of Special Education,* 1981, *15*(2), 239–256.

Page, E. B., & Jarjoura, D. Seeking the cause of correlations among mental abilities: Large twin analysis in a national testing program. Special issue on Intelligence, *Journal of Research and Development in Education,* 1979, *12*(2), 108–117.

Page, E. B., Jarjoura, D., & Konopka, C. Curriculum design through operations research. *American Educational Research Journal,* 1976, *13*(1), 31–49.

Page, E. B., & Keith, T. Z. Effect of U.S. private schools: A technical analysis of two recent claims. *Educational Researcher,* August 1981, *10*(7), 7–17.

Pedhazur, E. J. *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart & Winston, 1982.

Raiffa, H. *Decision analysis: Introductory lectures on choice under uncertainty.* Boston: Addison-Wesley, 1968.

Reschly, D. J. Psychological testing in educational classification and placement. *American Psychologist,* 1981, *36*(10), 1094–1102.

Scarr-Salapatek, S. Race, social class, and IQ. *Science,* 1971, *174,* 1285–1295.

Schmidt, F. L., & Hunter, J. E. Employment testing: Old theories and new research findings. *American Psychologist,* 1981, *36*(10), 1128–1137.

Stanley, J. C., & Hopkins, K. D. *Educational and psychological measurement and evaluation.* Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Taubman, P. (Ed.). *Kinometrics: Determinants of socioeconomic success within and between families.* Amsterdam: North-Holland, 1977.

Thompson, J. N., Jr., & Thoday, J. M. (Eds.). *Quantitative genetic variation.* New York: Academic, 1979.

Thorndike, E. L. The nature, purposes, and general methods of measurements of educational products. *The 17th Yearbook of the National Society for the Study of Education,* Part II. 1918.

Tillett, P. I. Optimization of secondary teacher assignments using operations research. *Socio-Economic Planning Sciences* (London), 1975, *9*, 101–104.

Trueman, R. E. *An introduction to quantitative methods for decision making.* New York: Holt, Rinehart & Winston, 1974.

VanDusseldorp, R. A., Richardson, D. W., & Foley, W. J. *Educational decision-making through operations research.* Boston: Allyn & Bacon, 1971.

Wagner, H. M. *Principles of operations research: With applications to managerial decisions.* Englewood Cliffs, N.J.: Prentice-Hall, 1969.

Winkler, R. L., & Hays, W. L. *Statistics: Probability, inference, and decision* (2nd ed.). New York: Holt, Rinehart & Winston, 1975.

Wright, S. Correlation and causation. *Journal of Agricultural Research,* 1921, *20*, 557–585.