DOCUMENT RESUME

ED 223 653 TM 820 746

AUTHOR Page, Ellis Batten
TITLE Rethinking the Principles of National Assessment: Towards a More Useful and Higher Quality Knowledge Base for Education.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Jul 82
NOTE 36p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Viewpoints (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Educational Assessment; Educational Principles; *Educational Quality; *Educational Research; Elementary Secondary Education; *Federal Programs; Political Issues; *Program Evaluation; Public Policy; Research Design; Research Utilization; Sampling; Social Problems
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

A review of the National Assessment of Educational Progress (NAEP), which originated 20 years ago, is presented; and the original political, ideological and social science assumptions for the design of NAEP are examined. The report criticizes the founding principle as hampering NAEP's maximum utility and exploratory power. In examining the form of NAEP, possible changes in design strategy and theory, sampling, measurements, reporting methods and administrative procedures are explored. Issues which emphasize the environment's influence on behavior in learning theory are discussed in relation to the uses of objective- versus norm-referenced testing. To benefit researchers, educators and public and private decision makers, a maximum amount of information should be collected from smaller numbers of students. Recommendations include a sounder scientific orientation, stress on curriculum-based learning, longitudinal studies, and frequent monitoring of NAEP. (CM)

# RETHINKING THE PRINCIPLES OF NATIONAL ASSESSMENT: TOWARDS A MORE USEFUL AND HIGHER QUALITY KNOWLEDGE BASE FOR EDUCATION

by

Ellis Batten Page
Duke University

Report commissioned by the National Institute of Education, U. S. Department of Education, as part of a restudy of the National Assessment of Educational Progress.

Duke University
Durham, NC 27708
July, 1982

# RETHINKING THE PRINCIPLES OF NATIONAL ASSESSMENT: TOWARDS A MORE USEFUL AND HIGHER QUALITY KNOWLEDGE BASE FOR EDUCATION

## INTRODUCTION

Compared with the U.S. total expenditure for educating our young, which has been called our largest national industry, the amount spent on educational research is pitiably small. All the more reason, then, that the present Federal Administration, committed as it is to reducing the budget wherever possible, should look closely at each major commitment in educational research, to appraise the past operations, with an eye toward improving the future, and of course to eliminating any programs which are not cost-effective.

The National Assessment of Educational Progress (NAEP) would, naturally, come under important review. It had its origin nearly 20 years ago, in a political and ideological (and social science) atmosphere quite different from today's. It has cost about 60 million Federal dollars, plus another 5 million from private funds (and huge uncounted expenditures in professional and student time at the state and

local levels). There are, then, very valid questions which we may raise: Are the operating assumptions of 20 years ago still the appropriate ones? Should it, indeed, survive in its present form? And if so, what changes should be made in its design strategy, its sampling, its measurements, its reporting, its administrative procedures, to improve its usefulness?

Other studies of NAEP have, of course, been conducted in the past. Some major studies have been reported by Greenbaum (1976), by the National Center for Education Statistics (NCES, 1974, 1975), by the General Accounting Office (GAO, 1976), by Wirtz and Lapointe (1982), and by Sebring and Boruch (1982). These have been useful in drawing attention to certain accomplishments and apparent inadequacies of the NAEP operation and dissemination. Most of these have taken NAEP's founding principles as a given, and some have lauded NAEP's methodological contributions to the collection of knowledge. A persistent plaint has been about NAEP's possible inadequacies as a useful and active aid to the state and local districts (SEAs and LEAs). These negative comments about low use have been skillfully parried by Sebring and Boruch (1982), who have pursued information or testimony from SEA and LEA personnel about NAEP usefulness, and from NAEP personnel about their activities at other educational levels.

In commissioning the present report (one of seven brief evaluations so commissioned), the National Institute of Education (NIE) explicitly asked:

> How can the NAEP design be enhanced to allow for the generation of information useful to:
>
> A. SEA and LEA and other state policymakers
>
> B. Federal policy makers
>
> C. Professional associations
>
> D. Research community
>
> E. The public.

This report will make a sharp break with the past evaluations of NAEP. I intend being much more critical of the founding principles underlying NAEP. I will argue that these principles were often rooted in the political and ideological commitments of the mid-1960's. Such a condition, if true, surely does not in itself invalidate the principles. But I shall further argue that these principles, however fashionable they may have been at the time, were at times crippling through the employment of poor sampling, measurement, and research design. Many of the more intellectual problems of NAEP (as distinct from the practical problems) apparently stem from such poor principles, themselves rooted in inadequate scientific understandings.

## NAEP'S IDEOLOGICAL UNDERPINNINGS

As most educators are aware, 1957 marked a strong shift toward federal support for education. The reasons, at that time, were not liberal but conservative: Russia had put into orbit its Sputnik space vehicle, and America, well aware of Russia's expansionist military drive and growing nuclear threat, saw education in a new light. Suddenly, there were grave national reasons for assuring U.S. parity or superiority in scientific and technical education relevant to military security. Indeed, the 1960 presidential campaign was marked by debate over the best ways to achieve such security, and the victor, John F. Kennedy, had charged that the former administration had permitted a "missile gap." In fact, the mood of Congress in the late 1950's was already strongly pro-defense and committed to strong federal involvement in education directed to this military aim. One result of such Congressional commitment was the National Defense Education Act. Another, particularly relevant to our brief history of NAEP, was the founding of Project TALENT.

The very name differences between TALENT and NAEP are revealing of the decades which gave them life: In one case, the nation sought the identification and nurturance of high-ability students (i.e., "talent") as an important national resource for the national strength. In the other, the nation considered not so much the talented individual, as it

sought political and legal ways of improving society as a whole. In Project TALENT, the most important goal was the recognition and cultivation of high individual intelligence. In the case of NAEP, intelligence hardly seems to have been an admitted reality (there are not even any explicit intelligence measures in NAEP's testing program). In TALENT, no ethnic information was asked; again, the individual was central. For NAEP, there was extensive information collected about race and SES, and other group variables, and the emphasis seemed to be more on issues of "equality," rather than of "excellence" (cf., John Gardner's influential book of 1961).

What had happened to shift the emphasis so markedly? Following the landslide election of L.B. Johnson in 1964, together with a liberal Congress, America moved toward an unprecedented federalization of social programs, marked especially by laws and regulations favoring groups regarded as having suffered discrimination in the past: Blacks, Latin-Americans in the U. S., Indians (who were renamed "Native Americans"), and women of all ethnic groups. University campuses, of course, are often the breeding ground for ideologies, and such was clearly the case during the mid-1960's. The spirit of revolt found student (and faculty) expression against the military draft, against perceived inequalities, and even -- perhaps especially -- against the campus "Establishment" itself. There was a widespread, and

very well-publicized, breaking down of former social restraints about public and private behavior, including social, sexual, sartorial, and political behavior. In general, reform was not expected from almost any traditional centers of authority, unless these were taken over and radically altered. One heard the expression, "Don't trust anyone over thirty!"

What was meant, then, by President Johnson's call for a "Great Society"? It was to be a society in which inequality was to be progressively reduced through the operation of the larger society itself -- and especially through that superordinate level, the Federal Government. There was an enormously increased level of federal taxation, of federal bureaucracy, of federal lawmaking affecting the conduct of almost everyone, and most certainly of the schools and colleges throughout the land.

Politics and Social Science

What we have been recalling, however, seems purely a question of political ideology, with little bearings on the structure of what should be, in principle, a scientific enterprise. After all, educational and psychological measurement and research have close kinship with statistics and mathematics: fields of an abstract, universal truth. And when these methods of science are applied to the observation of social and educational phenomena, then the empirical ob-

servations, and even their interpretations, should follow from such methods. We should be in a position to discover educational "truth," as well.

While such an ideal operates in principle, in the real world of the applied social sciences, the methods and beliefs are often servant to the overarching ideology, including most definitely the political ideology, and any other beliefs favorable to one's own well-being, or that of one's colleagues. In psychology, for example, a large number of one's colleagues earn a good living through the practice of "psychotherapy" -- despite the absence of any unified theory about its operation or of any persuasive evidence about its efficacy. With the exception of a few mavericks, often disparaged by their fellows, few eminent psychologists have seriously addressed this serious problem.

And the training of social scientists seldom includes systematic instruction in scientific ethics. Few textbooks of research methods make any mention of what one could hope would be the premier assumption of the scientific enterprise: That we attempt to discover only the truth, and that we expose and decry falsehood. Rather, in university trainings for advanced degrees in psychology, sociology, and similar fields, a higher virtue is sometimes regarded as the adoption of a liberal intellectual posture. Such a casual observation is indeed supported by careful research with

political questionnaires: Ladd and Lipset (1975) have repeatedly confirmed that the "social sciences," broadly defined, are the left-most major segment of academic faculty. Thus we reach the vexed problem of the relation between political ideology, on the one hand, and the conduct of scientific research, on the other.

## Environmentalism and Behaviorism

There is a complex of issues in which political ideology interacts with psychological theory. One of these key areas involves the relative emphasis on environment (vs. heredity) in its influence on behavior. In the major field of "learning theory," as taught during the 1950's and into the 1960's, human differences in learning were often treated (in the experimental literature) as an "error term" in the analysis of data. Indeed, to justify the huge theoretical leaps between the laboratory animals and the students in schools, it was common to leap even the differences between species. At the same time, of course, with the huge expansion in testing and the sudden access to computers, the field of psychometrics was looking exactly at these human differences, and making great strides. That these two fields, psychometrics and experimental psychology, had little to say to each other was pointed out by the educational psychologist Lee J. Cronbach, himself a psychometrician of note, in his Presidential Address of the American Psycholo-

sical Association, entitled "The Two Disciplines of Scientific Psychology." (His technical solution was that the analysis of covariance could combine human differences and experimental effects into a single statistical treatment: so much for the underlying theoretical clash!)

It is not surprising, then, that the swing of ideology would cause a corresponding swing in fashion among applied social scientists, from one of these emphases to the other. And in the activist 1960's, when many were looking toward government to resolve many human problems, a very persuasive spokesman for environmentalism was already present, in the noted experimentalist B. F. Skinner.

Radical environmentalism -- the emphasis on training to the exclusion of individual differences -- had of course an ancient and respectable tradition. John Locke's tabula rasa implied that the nervous system was a clean slate at birth (and of course one "clean slate" is like another). John B. Watson's vaunted claims for behavioral "conditionings" overrode any innate differences in talent or disposition, and assumed current mastery (even in the 1920's) of training technique. B. F. Skinner was the articulate inheritor of this behavioristic tradition and surely its most influential advocate in the history of psychology. To the joy of struggling graduate students, he relegated even learning theory to the ashheap of the unnecessary (Skinner, 1968). And with

his sweeping vision of virtually all behavior as under the control of what he termed "operant reinforcement," he generalized from rats in a Skinner box, to "Pigeons in a Pelican" (1971), to the productive and well-behaved fictional characters in Walden II (1948). Skinner's followers were legion, and as specialists in "behavior mod," they populated graduate departments in psychology and educational psychology with eager and apparently successful practitioners.

And their apparent successes ranged from the extinction of aggression, to learning the alphabet, to mastery of some school skills. Borrowing heavily from the methods of "task analysis" (e.g., Gagne, 1970), the behavior mod enthusiasts listed clearly the appropriate stimuli and the desired responses, elicited these responses from their students following the stimuli, and then rewarded the students (the reward was the "operant reinforcement" of the S-R pairing), and repeated the exercises with progressive "shaping" of the behavior toward the desired responses, which were the "criterion" or "objective" of the program.

Such environmentalism, the radical behaviorism that behavior mod represented, fit well into the government mood of the 1960's. If by changing the environment in this way, by providing intensive (often one-on-one) conditioning to the learners, and by using systems of enjoyable reward, we could effectively educate even low-achieving populations, then we might eliminate ignorance and poverty and misbehavior in one

generation. And the development of computers and special textbooks provided even an economical way to do all this, through "programmed instruction." But what was required, first and foremost, was to spell out clearly the stimulus-response pairings which would constitute the broadly desirable behaviors which were the objectives of the entire educational enterprise. If the Federal Government could help bring this about, then it surely should (an assumption loosely under the "equality of opportunity" mandate). And therefore the 1960's and 1970's saw a powerful surge of programs aimed at such "behavioral technology," and funded by federal agencies.

Unfortunately for the success of such an enterprise, however, its advocates were still not communicating with the measurement scientists. And when they did, they felt embarrassed by the irreconcilable nature of some of the underlying theories and practices from the two fields. Thus, for any persuasive rationale for behavior mod, it was necessary to develop a contradictory type of measurement theory. And this need was the powerful engine of the drive toward testings which would be the "criterion" or "objective" of the program.

## SELECTED ISSUES OF NAEP'S STRUCTURE

As we have seen, the ideology and political goals of the mid-1960's seemed to demand a set of social science conceptions which would rationalize these goals and permit their realization. A number of central issues of NAEP's operation, then, were resolved in compatible ways: NAEP would emphasize the importance of ienvironment, and diminish the importance of heredity. NAEP would emphasize the item-by-item learning of skills, rather than the (less easily altered) operation of general intelligence. NAEP would concentrate on the educational progress of groups, and give more importance to their equality than to individual talents.

To achieve worthwhile goals, it is not always necessary to employ sound theory. But in the recent decades of social science, poor theory has often led to results that were wasteful or worse. And several central issues from NAEP's history show weak theory which apparently has hampered NAEP in achieving maximum utility and explanatory power.

### Objective- vs. Norm-Referenced Testing

If behavior modification were to appear successful, then it must appear to achieve its stated objective: the change in behavior desired. If we are going to make claims of success for a new program of teaching reading, for example,

then we should be able to point to S-R pairings which have been clearly altered in the desired way. The trouble is, standard ability and achievement tests, unless they are themselves practiced during instruction, are not apt to be easily altered by behavior mod programs of reasonable length and expense. This difficulty was a severe one for applied behaviorism -- and there were in general just two solutions to it. One was to continue the use of standardized tests, but teach the items during the behavior mod instruction. The other was to argue for new instruments, specifically targeted during the instruction period, and to take these new instruments as the "objective" of the educational program. A third alternative was unacceptable: To use the standardized tests as criteria of the programs -- and to lose any refunding for those programs! Here are two historical examples, each originated in the 1960's:

Teaching the standardized intelligence test: The Milwaukee Project. The longing for a strikingly successful government intervention cannot be better illustrated than by the "Milwaukee Project," directed by Dr. Rick Heber of the University of Wisconsin-Madison. At the long-run cost of over $16 million in federal funds, Heber and his colleagues concentrated on just 20 disadvantaged children from a Milwaukee slum, and claimed I.Q. mean gains of 35 points, compared with their untreated controls. As it happens, despite the theoetical and symbolic importance of this demonstra-

tion, technical reports were fugitive and plagued by ambiguities. But the rare technical critiques of the Project reached the inevitable conclusion that the personnel had been systematically training the youngsters in the material of the standardized I.Q. tests employed to demonstrate success (Page, 1972b; Page & Grandon, 1980). (As it happens, Heber and one associate were later convicted of misuse of federal and state funds and are in the penitentiary, but the published critiques concentrated only on the scientific questions of the claimed effects.) Even with the expenditure of over half a million per child, the Milwaukee Project failed to produce measurable and generalizable, long-range effects in performance of standardized tests of fundamental importance, when these were not trained by the Project itself.

Teaching the standardized achievement test: Performance Contracting Experiment. There is no better test of behavior mod, broadly conceived, than a large experiment of unprecedented sweep, conducted by the Office of Economic Opportunity (OEO) and reported in 1971 (Page, 1972a). "Performance contracting" (PC) was a system of instruction through behavior modification, whereby students and psychologist-instructors were rewarded according to how well the students mastered instructional programs in reading and matheamtics. Ten thousand disadvantaged youths from three grade levels and from many different national sites were trained by en-

terprising R & D corporations, according to the leading theories of the behavior-mod school. All were optimistic about the outcome, on the basis of apparent successes of PC already demonstrated in Texarkana. Unfortunately, however, that earlier demonstration was marred by the discovery that students had been trained in the very items of the standardized tests used to evaluate the program. Hence, in this new large-scale OEO experiment, there would be no clouds of doubt: Trainers were forbidden to include material from the tests which would be used, under penalty of losing their pay. And the standardized tests were administered by rigorous evaluators unconnected with the R & D corporations and with the schools. The results were shattering -- so much so that the experiment is unknown to most students and practitioners of behavior mod: There was no discernible effect, at all, of the $6 million program. The teaching of the tests themselves appears to have been necessary to achieve the demonstrated effects.

NAEP's emphasis on objective-referenced testing. In light of this background, then, perhaps it is much more understandable why, in the conception of NAEP during this period of the mid-1960's, there should be so much argument in favor of objective-referenced testing (ORT) vs. norm-referenced testing (NRT). This shift in type of tests was presented to be one of the great contributions of NAEP, as com-

pared, say, with the Project TALENT which had preceded it. I remember large halls at national meetings of scientific societies such as AERA or APA, where Ralph Tyler and others argued forcefully about the advantages expected, if only the funding agencies came in for substantial amounts (and it was very clear that the important donor, following the pump-priming of Carnegie, would be the Federal Government).

This is no place to do a technical reanalysis of the differences, real or imagined, between the two kinds of testing. On one hand there are books such as James Popham's, making strong arguments for the uniqueness of ORT. On the other, there are the standard works on measurement and testing theory, such as Stanley & Hopkins' (19 ), putting these arguments in careful perspective, yet trying not to offend, at the same time, all the potential book-users who have believed the various arguments for ORT. Let us briefly consider some of these:

NRT is useful only for describing the differences between individuals. ORT, in this view, can really describe "what students can do," whereas NRT only says whether one student is better than another. This is simply not an accurate description of NRT, when it it is well done by competent testing agencies. Indeed, chapters in standard books on testing have long showed techniques of item analysis which would give the same sort of frequency descriptions for the tested

class, school, program, or other group of interest, as are claimed for ORT. On its part, ORT has no special way of describing what students can do. It is hard to imagine an item constructed for ORT which would not be just as appropriate in NRT -- in fact, a judge could not ordinarily tell which kind of test an item came from, if it were a standard, machine-scorable multiple-choice item.

NRT relies on multiple-choice items. In a view expressed for ORT, one could somehow get a clearer picture of "what students can do" if one broke away from the multiple-choice methods. And in NAEP's early days it was common to hear strong calls to professionals for new ideas, for innovative methods of testing. At NAEP, for example, writing samples were collected from huge numbers of students to find out -- never mind what the old normed tests said -- whether they could "really write." But then what? Then various educators and specialists were called in from across the nation to find out what to do with all these essays. There is, of course, no standard answer. When such essays are evaluated by experts in writing, their global (or holistic) assessments agree about .5 with each other, and contain far less information about specific student abilities and skills than one would derive from careful analysis of a well-designed standardized test on grammar, punctuation, spelling, and the like. (If a student does not use a quote in an essay, how will a judge decide his competence in use of quotes?) In

deed, NAEP has in practice depended greatly (and properly) on multiple-choice testing, with all its difficulties.

NRT is not responsive to program effects. As we have noted in the historical review, this charge is largely true. Most standardized tests are designed to measure something that is reasonably stable in the tested individual. The test is regarded as a kind of random sample of items from a virtually infinite population of items measuring the same ability. A short-term training program, therefore, would not be expected to alter, except marginally, this ability (whether we were considering an individual or a group of individuals). Therefore it is true that, for a highly specific training session (say, on the use of quotes in English writing), we can design test items to measure student acquisition of the objectives. This is commonly done by able teachers in the classroom. And most standardized tests in wide use will have at least a short sample of items on quotes. Can ORT do more than this? It is hard to see how, and NAEP's own record does not reduce the skepticism.

What we see, then, is that the emphasis on objective-referenced testing grew directly from the social drives of the period which gave birth to NAEP. Unfortunately, ORT has some rather large disadvantages, which had better be noted here as well:

ORT often has fragmented objectives. In NRT, the theory of ability testing is fully developed and very useful. We can easily calculate means, standard deviations, and correlations among abilities. We can use techniques to estimate the ranges of the true scores, underlying the observed scores. All of normal-curve theory comes into play. We can describe contrasts between groups and programs in well-understood ways (for example, in terms of means or, more powerfully, of median overlap). In contrast with such uses of NRT, ORT is peculiarly hamstrung by its need to justify its test philosophy. NAEP's reports, therefore, often consist of item-by-item response frequencies for groups, and the larger, more efficient comparisons are lost sight of.

ORT is not statistically tractable. This is no light charge, for the truth of "national progress," once one steps away from the straitjacket of objective testing, must rely heavily on advanced statistical methods. The tracking of change across time (for example in study of the test-score decline) must depend on having the best statistical tools at our disposal; and there is nothing in ORT to compare with the brilliant normal-curve theory we have inherited from De-Moivre, Galton, Spearman, and our other progenitors.

ORT is awkward in causal study. After all, the purpose of NAEP is to provide information for national decision-making. And decisions depend directly on information about the

causes of student achievement. Apart from experimentation (which is virtually impossible to conduct efficiently in the schools today), our best methods for exploring causation currently come from path analysis -- the collection of methods for estimating the influences of variables on each other (cf., Kenny, 1979; Heise, 1975; and many others). All path analytic methods begin with correlations or covariances, which depend on having scores which permit the expression of such relationships. I know of hardly any causal studies, using such powerful methods, to have come out of NAEP (in striking contrast, for example, with the work done with Project TALENT, or with the more recent National Longitudinal Study, or with High School and Beyond). This is a most serious limitation of the objective-referenced framework -- a limitation, moreover, striking at the heart of NAEP's founding purpose.

Sampling Issues of NAEP

Just as NAEP's measurement theory grew out of the 1960's and the behaviorist movement, so did NAEP's sampling strategies reflect the same spirit. If achievements are a collection of specific skills, to be trained and tested, then it is less important to know much about the individual differences of the learners. What is more important is to measure those specific skills. The principles of NAEP included a severe limitation (one hour) on the time required of any one

student. Rather, different items and measures were applied to different groups of students -- with only a skeleton of background information on each participant. In the same spirit, there was no effort to look at the developmental variables best studied by repeated investigations of the same students across time. In its sampling structure, once again, NAEP has resembled the behavior modification group of psychologists relatively more than those in measurement or child development.

From the standpoint of research value, the loss of information from the NAEP sampling strategy is incalculable. The information from very short vectors of data (with few items on each student) is slight -- and it is no wonder that there has not been very much serious research use made of NAEP's data, compared with smaller and less expensive data sets of a more measurement-oriented design.

What is often not realized is that information is roughly proportional, not simply to the number of information items on each subject, but proportional to the square of the number of items. This happens because available information may be thought of as the number of correlation coefficients possible from a dataset. The actual formula is that:

$$C = n(n-1)/2, \qquad \text{Equation (1)}$$

where n is the number of items on each student, and where C is the number of bivariate correlations obtainable from the matrix. This is not to mention the much larger number of three-variable and n-variable complex relationships expressible from longer data vectors. So it is clear that all those interested in research -- notably university researchers and scientific societies -- are losers by the sampling strategy adopted by NAEP.

What is not clear to many is that virtually all potential users of NAEP's output are similarly deprived. As we have noted, most of our information needs are concerned with causes, not simply with observed data. And decision makers wish to know the probable effects resulting from their choices. Simple-minded one variable and two-variable data reports are not apt to meet anyone's needs, for they are often as misleading, from the standpoint of decision making, as they are informative.

The principal message of this portion, then, is that NAEP's sampling design, like its measurement theory, needs to be rethought from the ground up, and freed of some of the doctrines implicit in its founding.

## The Problem of Access to NAEP Data

Obviously, data are not valuable which are not analyzed. What is less obvious is that, given a truly rich data set,

it is impossible to predict all of the useful analyses. To the contrary, the serendipitous discoveries from data can, at times, bear fruit surpassing the original intention. And, depending on who is doing the analysis and on what assumptions are made, different investigations can result in strikingly different conclusions about the results, and hence about the appropriate decisions to recommend. (E.g., by using different controls Pase & Keith, 1981, produced very different comparisons of U.S. public and private high schools from those drawn by Coleman et al., 1981.)

In the old, experimental models of the psychological past, the data were seldom available to outside researchers. And even the monumental Project TALENT did not produce data tapes easily available to researchers of different purposes or views. For the educational research community, the first breakthrough of access has been with the National Longitudinal Study (NLS) of the High School Senior Class of 1972. The NLS has operated exactly from the viewpoint expressed here: that good data sets are inexhaustible and may be mined richly, and indefinitely, by a host of researchers for whatever purpose. Thus, the NLS (sponsored and monitored by the NCES) has emphasized the availability of data much more than its analysis. With encouragement to researchers, with low costs, and with widespread information about the NLS, together with the base year (1972), and four follow-up years (the latest being 1980), for a total data vector of some

2,000 measures, the NLS has attracted great interest and use.

The NLS's direct successor has been the still richer High School and Beyond (HSB), which studied 58,000 high school seniors and sophomores in 1980, and has already put in motion its first follow-up (1982). I have argued elsewhere that HSB may be the single best event that ever happened for educational and psychological research (Pace, 1981). Among other innovations, it is a great leveler: Anyone can get the data for low cost. A brilliant graduate student may glean more riches from it, in basic or applied educational knowledge, than a distinguished professor or head of a large institute. It permits the original analysts, together with their critics, to compete on equal terrain, their arms consisting in their ability to think, analyze, and report. Journal editors are clearly hungry for the kind of generalizability which such data sets make possible.

And decision makers, from state and local and federal agencies, can explore the data with their particular questions in mind. These will very often be different from the routine information given out by any one general agency. Of course, it remains a problem to establish skillful data analysts to help decision makers, but such analysis may be drawn from skilled researchers in universities, or in other R & D agencies. It is a knowledge contribution which cannot

be overemphasized, made by the openins of such rich data to universal scrutiny. It is also the mark of a remarkably confident and open form of sovernment that a nation is willins to do so.

In lisht of these developments with the NLS and with HSB, may the NAEP archives be made equally accessible? One of NAEP's most valuable recent chanses is in the plan to produce widely useful public data tapes. Already some of these are in place, and one may expect many more in the future, siven the strons intention, within the NIE, to achieve this soal. These steps are to be sreatly encourased; many of these data sets will be unique for some major lines of investisation (for example, the promised tapes from the writins samples, in machine-readable form).

Nonetheless, one liability should be pointed out affectins NAEP far more than these other data sets. As we have noted, NAEP's use of many different student samples means that researchins complex relations will be much more difficult, because the coefficient of information C in Equation (1), noted above, will be a much smaller fisure. In future versions of NAEP, it is hoped, a maximum amount of information will be collected from smaller numbers of students. Then the sreater access to NAEP data will be of much larser benefit for researchers, for SEAs and LEAs, and for all those decision makers, public and private, with a stake in American education.

## RECOMMENDATIONS FOR NAEP'S FUTURE

From the present analysis, and from other considerations as well, there are a number of recommendations which may enhance NAEP's future utility, for all of the targets of its concern: federal, state, and local policy makers, professional associations, the research community, teachers, students, and the public. These recommendations will be here made under the headings of design strategy, sampling, measurement, reporting, dissemination, and administration.

### Design Strategy

1. Look for all explanatory causes of important student behaviors. Include influences of the family (both environmental and biological), of the church, neighborhood, and other major influences outside the school.
2. Continue to stress curriculum in its relation to learning. Include not only courses taken, but homework required and homework actually done.
3. Save design money by using the lower-cost multiple-choice items where appropriate, including those from instruments already available and on the market.

4. Also save costs through planning the use of fewer student subjects with more contact time per student. (These points are repeated below.)

Sampling Issues

1. Abandon most of the separate student samples. Move toward more examination of fewer students to create the most informative data matrix possible.
2. Examine students longitudinally, showing educational and other growth across the school years. Keep the emphasis on the school years, however, and leave to other programs any adult follow-up.
3. Continue matrix sampling of achievement, where appropriate and necessary for economy and reasonable data burden. But use the economical standardized tests wherever justified.
4. In the sampling of students, include a good number of students of high ability and achievement (as well as other target groups) since these are important to national productivity and other national interests.
5. Include a large sample of siblings and twins of the student subjects, for better explanation of home and of school effects.

Measurement Issues

1. As noted, available instruments, both from standardized programs and from other researches, would be economical replacement for many items created by

NAEP, and should be used. Continue to monitor educator and lay opinion about the objectives of education. Use these findings to help structure inquiry into skills and curriculum, and to validate the tests chosen for use.

2. At the same time, recognize the essential sampling nature of all achievement testing, the heavy loading of general ability in such measures, and their normal distribution. For each student, report scores which may be used in "norm-referenced" ways for explanation and prediction.

3. Measure a range of different sub-tests of ability, including verbal and nonverbal items of intelligence. Use questionnaire items for extensive information about social, familial, ethnic, and physical and health characteristics of the students. Again, these may include standardized or other available instruments where appropriate.

4. For twins of the same sex, determine their kinship (whether identical or fraternal). Such information can help in causal explanations of school achievements, and of other behavior.

Reporting Results

1. Continue to mail out descriptive material about the studies and their most general results.

2. In addition, however, solicit some advanced causal analysis of key issues, to attract the attention and use of researchers and policy makers. Welcome different analyses and debate in the public forum.
3. For the SEAs and LEAs, report information on current changes discovered by the latest surveys, in student attitudes and achievements, and in school organization and curriculum.
4. In reporting, keep in mind that education seeks general liftings of student abilities, that items are only samples from these larger traits and achievements.
5. In group comparisons, use more informative indices than "percentage correct." For continuous traits, use such comparisons as median overlap, or simple phrases carrying the same meaning.
6. In general, abandon the language of passing or failing. There is no such definable line in the world of normally distributed, latent abilities.
7. Where possible and reasonably consensual, show the causes of differences noted. In practice, this may mean approximated by showing differences after controlling for background variables.

Dissemination of Data

1. Follow through with present plans for clean, well-documented data tapes available at very low cost to any requesting users.

2. Organize these tapes around the large body of extensively researched students described above. That is, in the recommended future structure, these will not be usually organized around separate curricular issues

3. For those different data segments already collected, organize materials with extensive description which may be widely announced to prospective user groups.

Administration of NAEP

1. Continue advisory groups from informed professional and lay constituencies, but also experienced researchers and measurement personnel of the highest quality. This should make for a sounder scientific orientation in the NAEP operation.

2. In the scientific advisory role, include more experts experienced in and dedicated to causal analysis and to the norm-referenced tradition.

3. For economy and efficiency, put the NAEP contract periodically out to rebid. As recent experience shows elsewhere, competitive bidding may shave more than a million dollars a year from the annual cost of programs of this magnitude.

In summary: As thinkers have recognized since Plato, there is no more important task of government than education. Its process, and progress, must be frequently monitored by society at large. We now possess improving methods for making such monitoring effective, and for interpreting the results in meaningful and useful ways. When organized well, information systems may provide both simple and complex analysis to deepen our general understandings of the educational process. And through the use of simpler language, these more sophisticated interpretations can be expressed to the targets of all our concern: the SEAs, LEAs, schools, educators, parents and children and citizens -- to all who wish to improve the quality of human life and human learning.

## REFERENCES


Coleman, J., Hoffer, T., & Kilsore, S. Public and private schools: A report to the National Center for Education Statistics. University of Chicago, National opinion Research Center, March 1981.

Gagne, R.M. The conditions of learning, 2nd ed. New York: Holt, Rinehart & Winston, 1970.

Gardner, J. Excellence, 1961.

General Accounting Office. The National Assessment of Educational Progress. Its results need to be made more useful. Washington, D.C.: National Center for Education Statistics, July 20, 1976.

Greenbaum, W. Measuring educational progress. New York: McGraw-Hill, 1976.

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

Kenny, D. Correlation and causality. New York: Wiley-Interscience, 1979.

Ladd, E. C. Jr., & Lipset, S. M. The divided academy: Professors and politics. 1975.

Pase, E. B. How we all failed in performance contracting. Educational Psychologist, 1972a, 9(3), 40-42.

Pase, E. B. Miracle in Milwaukee: Raising the IQ. Educational Researcher, 1972b, 1(10), 8-16.

Pase, E. B. The media, technical analysis, and the data feast, A response to Coleman. Educational Researcher, August 1981, 21-23.

Pase, E. B., & Grandon, G. M. Massive intervention and child intelligence: The Milwaukee Project in critical perspective. Journal of Special Education,

Pase, E. B., & Keith, T. Z. Effects of U.S. private schools: A technical analysis of two recent claims. Educational Researcher, August 1981, 7-17.

Pase, E.B., & Pihanso, R. Information gained through combining psychometric studies. Proceedings of the American Psychological Association, 1970, 127-128.

Sebring, P. A., & Boruch, R. F. On the uses of the National Assessment of Educational Progress. Division of Methodology and Evaluation Research, Northwestern University, Evanston, April, 1982.

Skinner, B. F. Walden II. 1948.

Skinner, B. F. The technology of teaching. New York: Appleton-Century-Crofts, 1968.

Skinner, B. F. Pigeons in a Pelican. Gold Medal Award Address, American Psychological Association, 1971.

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

Wirtz, W., & Lapointe, A. Measuring the quality of education: A report on assessing educational progress. Washington, D.C.: Authors, 1982.