DOCUMENT RESUME

ED 088 898 TM 003 407

AUTHOR	Mason, Ward S.
TITLE	Problems of Measurement and the NIE Program.
INSTITUTION	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	29 Aug 73
NOTE	79p.
EDRS PRICE	MF-$0.75 HC-$4.20
DESCRIPTORS	Evaluation; *Evaluation Needs; Measurement; *Measurement Techniques; Program Development; *Program Planning; Program Proposals; Testing; *Testing Problems
IDENTIFIERS	*National Institute of Education; NIE; NIE Archives

ABSTRACT

A need exists for the National Institute of Education (NIE) to extend the range of its concern with measurement into a number of new areas. While the measurement of basic cognitive abilities is well-advanced, accurate measures of affective and higher-order cognitive abilities are not generally available. Measurement could also be extended into other dimensions as well; specifically, the advancement of the ability to measure systems; the development of the measurement sub-disciplines of sociology and political science; improvement of unobtrusive data collection methods such as observation; better support for the research and development community; detection and measurement of unplanned consequences of educational programs; identification of inputs, contexts, and processes related to educational outcomes; emphasizing the importance of theory in deciding what needs to be measured. The author presents tentative recommendations for initiatives into the newly-defined areas of educational measurement. (NE)

TM

PROBLEMS OF MEASUREMENT AND THE NIE PROGRAM

Ward S. Mason
National Institute of Education
August 29, 1973

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

A number of colleagues were kind enough to offer helpful comments on an earlier draft of this paper. These include Marvin C. Alkin, Robert B. Binswanger, Ronald G. Corwin, Lois-ellin Datta, William Dorfman, Abbott L. Ferriss, Willis W. Harmon, Robert E. Herriott, Beverly Kooi, William G. Spady, Thomas C. Thomas, and David V. Tiedman. I have also benefited from discussions with and working papers prepared by the participants in the October, 1972 conference sponsored by the Center for the Study of Evaluation and the Educational Testing Services, listed on page 3. Nancy Holt was assiduous in tracking down source material and in providing editorial assistance.

Of course I take full responsibility for any shortcomings in the final product.

PROBLEMS OF MEASUREMENT AND THE NIE PROGRAM

Introduction

Background

Attention to problems of measurement has been a salient concern since the first thinking about the National Institute of Education began. Indeed, the President's message on educational reform, which first placed the formation of NIE on the government agenda, highlighted the need "to develop broader and more sensitive measurements of learning than we now have" (Nixon, 1970). This need was placed in the context of the need for accountability of schools and teachers so that our educational institutions might be more responsive to local requirements.

The establishment of an NIE Planning Unit inaugurated an extensive planning process. Prominent individuals and groups of experts prepared a wide assortment of papers, some focused on the contributions which various disciplines might make to the study of education, some focused on specific educational problems, and some providing syntheses of specific recommendations (NIE, OPI 1973). An analysis of these papers revealed that the need for new measures in education was a common theme running through many of them (Kooi, 1972).

> Writers of the NIE planning documents agreed that new measurement procedures could be the basis for changes in the present structure of education and allocation of resources within it, or measures could provide new bases for credentialing so that current educational requirements could become more flexible. However, a program of exploration and development would be needed to realize this potential. Though there are some widely used tests that might adequately assess proficiency

> in reading, mathematics, and the sciences, there are virtually no generally acceptable instruments for assessing complex problem-solving skills and social-emotional behavior. For NIE to sponsor development of even rough milestone measures of learning in these domains would represent a vital and useful beginning. The purpose of this NIE initiative is to take the first step of examining educational measurement needs and designing a program to fill gaps in the area. During the coming year, the Institute should explore new techniques such as criterion-referenced (or domain-referenced) tests which sample behaviors and skills in specific areas directly and do not attempt to compare the student with others nor to predict his future ability. Another promising direction--both for individual measures and for developing social indicators for learning situations--ties in the expansion of direct observational methods.
>
> Before new techniques are expanded, however, the availability and sufficiency of measurements must be determined. Information is needed on what behavior should be tested, what tests are available, and how current measurements will work. When promising measures are identified, but validity, reliability, or standardization data are missing for them, this data should be collected. Such a study will identify gaps in traditional and new measurement so that a rational NIE program can be designed.

The crucial need for the improvement of measurement in the disciplines underlying educational research has also been expressed. For example, the prominent sociologist/methodologist, Hubert M. Blalock notes that:

> ...certain kinds of inadequacies in measurement procedures may very well provide the major obstacle to be overcome if sociology is to mature in the direction of becoming a "hard" and disciplined social science. (Blalock, 1969)

The Institute was actually established in 1972. The authorizing legislation lists four purposes for NIE:

- help to solve or to alleviate the problems of, and promote the reform and renewal of American education;
- advance the practice of education, as an art, science, and profession;
- strengthen the scientific and technological foundations of education; and

- build an effective educational research and development system. (Education Amendments of 1972, Title III, Sec. 405. (a)(2), p. 99.)

The need for good measurement is basic to all these objectives, but perhaps it is most convenient to think about it in relation to the third and fourth objectives. Good measurement is part of virtually all educational processes, beginning with the teacher's need to assess the performance of her pupils and including the assessment of teachers and schools, and making decisions and resource allocations at local, state and national levels. Because measurement is so basic, it will be inevitably a part of any program which NIE undertakes. One of the issues which this paper must consider is which measurement-related activities are most appropriately organized on a focused, centralized basis and which are best handled within the context of specific programs.

With the formal establishment of the Institute, new measures in education was recognized as the subject for continued program development work, first within the context of the New Initiatives Task Force and then as part of the Exploratory Studies Unit. A small conference was held in Princeton on October 2, 1972, under the sponsorship of the Educational Testing Service and the Center for the Study of Education.*

*Conference participants were: Scarvia B. Anderson, Samuel Ball, Samuel Messick, Elsa Rosenthal, and E. Belvin Williams, all of ETS; Cornelius Butler and Ward Mason, both of NIE; Donald Fiske of the University of Chicago; Douglas Jackson of the University of Western Ontario; Silvan Tomkins of Rutgers University; Stephen Klein, Beverly Kooi, and Robert Pace of CSE.

Following the conference, two documents were prepared. Beverly Kooi, a consultant to NIE, drafted a statement summarizing the statements prepared by conference participants in eight problem areas thought of as a system of interacting variables (Kooi, 1972):

- Personal and social values and their educational implications
- Treatments as experienced by individual learners
- General environments in which learning takes place (including home, community, and school)
- Specific aspects of cognitive/intellectual development
- Specific aspects of personal/social development
- Cognitive styles
- Theory and methodology (evaluation and research design; methodology of measurement per se and of research design)
- Costs (people and financial)

Second, Mason outlined some tentative program ideas for NIE derived from the conference results organized around two themes: (a) activities aimed at building the R&D infrastructure, and (b) activities aimed at collecting and analyzing data for use in policy research. (Mason, 1972).

It is the purpose of the present paper (1) to provide a broad survey of issues and problems in education and educational R&D which relate to measurement; (2) to present an overview of current NIE activities which are relevant to these problems and issues, and (3) to present some tentative recommendations for NIE initiatives. The recommendations are tentative for several reasons. The scope of this field is so broad that it would be impractical to present a thorough

analysis of each problem leading to a final recommendation; nor would any one individual be competent enough to make an equally credible presentation of every issue. Further, it is important that the staff assigned to develop any given program have a central role in developing specific program plans. It is hoped that this paper will be able to identify some "places to start", and that appropriate organizational units or task forces can be formed to refine, elaborate or reject each recommendation, as may be most appropriate.

What Needs to be Measured?

Although much of the discussion of the need for new measures in education has focused on the needs to measure pupil outcomes other than the usual cognitive skills, this is only part of the problem. Herriott and Muse make the useful distinction between variables at the individual level and those at the system level and note that such variables can serve as either independent or dependent variables (Herriott and Muse, 1973). A cross classification of these elements produces the following typology:

Classificatory Schema Depicting Focus of the Independent and Dependent Variables in Studies of Educational Effects

Dependent Variable	Independent Variable: Individual	Independent Variable: System
Individual	A	B
System	C	D

They point out that most educational research traditions can be classified in one of the cells of this table. Thus much of the research in educational psychology seeks to relate the personal and behavioral characteristics of teachers to test scores of pupils (cell A); social psychology has fostered a line of inquiry focusing on the impact of institutional factors on students, mostly at the college level (cell B); and economics has confined itself largely to the study of production functions of education - how educational resources interact with student characteristics to produce variation in student behavior (cell D). They point out the limitations of each of these traditions and call for the development of more comprehensive conceptual frameworks.

A key point is that a given variable can play various roles, depending on the problem and the analytic scheme. Thus a measure of student attitudes might be important both as an input and an output variable; if the same variable were aggregated by peer groups it might be a measure producing contextual effects. Thus it is not possible to classify measures in terms of their analytic role; NIE needs to be concerned with the development of measures serving many analytic functions, and not simply pupil outcomes.

Who is the Client and What is the Purpose?

It is a generally accepted principle that somewhat different kinds of measures have to be constructed for different purposes. Cronbach distinguishes (a) selection and classification of persons, (b) evaluation of treatments, and (c) checking on scientific hypotheses (Cronbach, 1970).

Identifying different clients or users also helps to identify different purposes. Practitioners are the primary clients of the testing industry. Traditional uses of test information by teachers include diagnostic and prescriptive decisions regarding individuals and groups. Administrators use test information for making decisions regarding programs and allocation of resources. They rely on many other kinds of data as well. Student record files contain information on pupil achievement, plus health, family and other kinds of data. Schools and school systems also have elaborate record keeping systems for fiscal, personnel, and other information which provide statistics for local, state, and federal use. Increasingly these various kinds of data are being used for program evaluation and as parts of management systems seeking to assure "accountability".

The researcher generally has rather different purposes in mind. Primarily he is interested in relationships among variables and in making causal inferences. Researchers can make a great contribution to the determination of the construct validity of measures by showing how they are part of systems of variables, and through studies of the population and ecological validity of measures, showing what variations in interpretation follow from the use of given measures with different sub-groups and in different contexts. (Anderson, Messick and Hartshorne, 1972; Cronbach, 1971).

The developer generally has purposes that overlap those of both practitioners and researchers. To the extent that the product to be developed incorporates the use of tests or other measures, the developers

purposes coincide with those of the practitioner. However, the development process itself requires the use of measures for purposes like those of researchers and evaluators.

The special needs and perspectives of evaluators, policy makers, change agents, and others might also be detailed. NIE needs to concern itself with this total range of clients and purposes, and not simply with the development of new tests for use in operating school systems.

Measurement of Individuals

New Learning Outcomes.

The most common point of entry into this problem area has been the observation that education has been focused on too narrow a range of cognitive outcomes and that measures should be developed for other kinds of objectives. This is, of course, in the first instance an argument concerning the goals of education rather than measurement per se, but implicit is the thought that we often pay more attention to things that have been quantified. For example, the President's message on educational reform called for new measures of achievement:

> To achieve...fundamental reform, it will be necessary to develop broader and more sensitive measurements of learning than we now have.
>
> The National Institute of Education would take the lead in developing these new measurements of educational output. In doing so, it should pay as much heed to what are called "immeasurables" of schooling (largely because no one has yet learned to measure them) such as responsibility, wit, and humanity as it does to verbal and mathematics achievement.
>
> (Nixon, 1970, p.3)

In a report prepared for the NIE Planning Unit, Etzioni distinguishes between instrumental and expressive goals of education and states that we have tended to overemphasize the instrumental goals. He feels that this imbalance should be corrected and calls for the development of expressive tests (Etzioni, 1972).

Another planning report, by Kooi and Associates, provides a goal analysis structure as follows (Kooi et. al., 1972):

A. Learning goals

1. Social and emotional development
   a. Self-acceptance
   b. Relating to others
   c. Responsibility
   d. Adaptability

2. Cognitive development

3. Physical development

B. Enabling goals

C. Systems Goals

1. Productivity

2. Access

3. Participation

This schema is especially useful in making it clear that not all educational goals can be reduced to learning goals. Each goal area implies need for measures to assess progress toward the goal.

Levien also calls for:

> ...development of techniques and instruments for evaluating a far broader range of education results than are commonly considered. Among the requirements are:

- Methods for assessing psychological development, cognitive and motivational...
- Methods for assessing learning outcomes referenced to objectives...
- Methods for assessing social development...
- Methods for assessing the development of learning skills and incentives.

Techniques should also be developed for identifying and measuring some of the reasonably objective consequences of educational programs on society, and some of the educational effects of outside-the-school influences--family, friends, television (Levien, 1971, pp. 79-80).

Krathwohl and Payne note that educational objectives for individuals can be stated at three or four levels of specificity. (Krathwohl and Payne, 1972,). At the most general level, there are many statements or objectives that have been formulated by national commissions, professional groups, and prominent individuals. Such statements commonly give as much prominence to non-cognitive objectives as to cognitive. However, they note that in curriculum building efforts complex objectives are likely to drop out.

> This erosion-of-effort is particularly likely to occur with affective objectives. The conceptual structure of nearly all new efforts at curriculum building includes affective objectives in some important way. But as the structure is developed, such objectives cease to influence the direction of instruction, the choice of activities, or what students learn. As objectives to be achieved concommitantly with cognitive objectives, they are <u>not</u> taught directly, and it is often merely hoped that they will be achieved with not concentrated effort on them...

> An additional important factor is that students will typically seek to learn those aspects of a course that will earn them a good grade, and affective objectives rarely play any significant part in grading. (Krathwohl and Payne, 1972, pp. 35-36.)

However, there is some question concerning the degree to which one should expect affective objectives to be reflected in and achieved through the explicit curriculum system. There are many elements of social structure and process which in effect constitute an implicit curriculum having important consequences for the affective outcomes of education. We also need to note the importance of many other factors such as family values and community contexts in determining affective states. The point is not to question the importance of measuring affective outcomes but to question the apparent assumptions that all such outcomes need to be represented in the explicit curriculum or that they are solely determined by school experiences. Given such multi-factor determination, if we are to measure affective outcomes we must avoid simplistic models which ascribe to the schools the sole responsibility for determining such outcomes.

The determination of what new learning outcomes need to be measured is, of course, partly a matter of the selection of goals and objectives, and is thus a political process requiring input from many sectors. NIE is supporting several activities which contribute to this process.

- The Center for the Study of Evaluation at UCLA has developed a needs assessment kit which provides a means for local schools to work with community members to identify and select goals for school programs.

- The NIE Research Division is conducting a series of surveys and laboratory studies to map the educational goal structure of the lay public.

- The Office of Research Grants is supporting a follow-up study of Project TALENT participants which will assess the efficacy of past and present educational programs which ostensibly prepare individuals to achieve their life goals and is expected to contribute to the formulation of educational priorities and goals.

However, not all needed outcome measures can be tied to pre-determined objectives. Sociologists have long stressed the importance of searching for unintended and unanticipated consequences of purposive social action (Clark, 1973), and this point has been emphasized by Michael Scriven in the context of educational evaluation (1972). Clearly we need to be able to detect and assess effects whether or not the program designer planned them. Sensitivity to possible side-effects might come from use of a different disciplinary perspective, or from insight born of experience.

Cronbach has expressed the dilemmas about whether what we can measure are the most important things, and whether to emphasize the empirical or theoretical approach to instrument development:

> Only the strict empiricists, those who eschew theory as entanglement, have been marketing practical new products and procedures. I cannot escape the feeling that the things actuarially scored tests cannot do are more important than the things they can do. Is the time not ripe for a wholly fresh effort to construct a new generation of tests? Or must testing based on theory wait until theoretic and metatheoric problems are better resolved? (Cronbach, 1970, p. xxviii).

From both the R&D perspective and that of practical use, conceptualization is of great importance in identifying new measures.

There are several advantages.

> Measurement development pursued as part of a theoretical framework instead of on an ad hoc basis permits one to (a) evaluate the adequacy of the measurement in terms of the meaning of the construct, (b) consider individual score differences as representing more or less the trait measured, and (c) compare and integrate results across studies in terms of common constructs.
>
> If we eventually want to use measurement for practical purposes such as diagnosis and evaluation, we must be prepared to justify that use in terms of the social consequences, and these cannot be evaluated without information about the meaning of the measure. No accumulation of sterile statistics can compensate for lack of understanding. (Anderson, Messick, and Hartshorne, 1972, p. 2).

It is not possible within the confines of a paper like this to come up with a specific list of variables for which new measures are needed. What we urge is that program managers and evaluators throughout NIE become sensitized to the need to consider a much broader range of human abilities (as both inputs and outputs). This is already going on within a number of programs. However, there is a problem in that these efforts tend to go on in isolation from one another; for example, there is a lack of compatibility among the measures used for the evaluation of different Career Education models and among those used by the several evaluation contractors of the Experimental Schools Program. In the final section of this paper an agency-wide task force is recommended which would help to identify common needs for new measures among programs and coordinate measurement development activities. Not only would possible redundancy be avoided, but an important contribution would be made to forming bridges of comparability among programs. One

of the most important barriers to the cumulation of knowledge in educational research has been the lack of agreed upon common measures of educational phenomena. To the extent that NIE can provide leadership in identifying and developing new measures of wide use and credibility, it will have taken a major step toward improving the cumulative character of the knowledge base.

While current problem-oriented programs are providing some support for measurement development, it is the nature of the case that these efforts tend to be short range and program dependent. Furthermore, it is difficult to put aside sufficient program money for measurement development when the thrust of events is to "get the job done". As part of the matrix management scheme proposed below it is therefore recommended that the agency-wide Task Force on Measurement have funds at its disposal with which to support the development of new measures which are expected to be of wide applicability in research and/or practice.

RECOMMENDATION:

- The NIE budget should set aside $300,000 in FY 74 and $1,000,000 in FY 75 for development of new measures of wide applicability in research and/or practice. These funds should be under the control of the agency-wide Task Force on Measurement and would be supplemental to funds used by individual programs to develop program-related measures.

So far the discussion has been confined to the measurement of pupil outcomes. In the last few years the measurement of teacher competencies has achieved considerable importance with the passage of legislation in several states requiring that teachers be evaluated on their competencies (Popham, 1972). Although the history of research on "teacher effectiveness" is long, its results have been meager. The new legislation found the field quite unprepared with regard to the availability of a suitable array of teacher measures.

The Office of Research and Exploratory Studies has a Task Force on Education Personnel. The role of the teacher is of crucial importance in any educational program, and the work of this unit has the potential of considerable impact on other activities within NIE. Improvement in the conceptualization of teacher functions and their measurement should play a central part in the work of this unit.

RECOMMENDATION:

- The Task Force on Education personnel should give a prominant place in its program to the development of measures of teacher competencies and activities as needed for new teacher accountability regulations and and for the implementation of innovative educational programs.

Availability and Quality.

Some empirical data have been published on the relative importance of different educational goals and on the availability of tests for the different goals. In its continuing program of evaluation technologies the UCLA Center for the Study of Evaluation obtained data from a national sample of 2,555 elementary school principals, teachers, and parents on their ratings of 106 educational objectives. Although most of the objectives listed referred to cognitive skills and knowledge in a variety of subject areas, the ten top-rated goals were mostly non-cognitive.

Top Ten Goals for Elementary Education Derived from Ratings of a National Sample of Principals, Teachers and Parents, and Availability of Published Tests for these Goals in 1970-71.

Rank	Goal	No. of Tests*
1	Self-Esteem	5
2	Citizenship	0
3	Socialization-Rebelliousness	11
4	Need Achievement	1
5	School Orientation	9
6	Neuroticism-Adjustment	30
7	Listening Reaction and Response	15
8	Attitude Toward Reading	0
9	Silent Reading Efficiency	21
10	Dependence-Independence	16

Source: Hoepfner, Bradley, Klein, and Alkin, 1972, p. 24; and Hoepfner, in press.

The availability of tests is very uneven for both cognitive and affective objectives. For all 106 goals, the correlation between the rating of importance and the number of tests available was only +.27, and

many goals had no tests at all. (Hoepfner, in press.)

The availability of a test and its quality are quite separable. The major sources of information about test quality are the Buros Handbooks and the Test Evaluations published by the Center for the Study of Evaluation (the latter with NIE support), although some of the other compilations cited in the bibliography have such information. Both Buros and CSE state that many tests are of relatively low quality.

> Test publishers continue to market tests which do not begin to meet the standards of the rank and file of (Mental Measurements Yearbook) and journal reviewers. At least half of the tests currently on the market should never have been published. Exaggerated, false, or unsubstantiated claims are the rule rather than the exception. Test users are becoming more discriminating, but not nearly fast enough. (Buros, 1972, pp. xxvi-xxviii)

And CSE, commenting on its evaluation of tests of higher order cognitive affective, and interpersonal skills:

> In conclusion, it should be noted that, in the opinion of the CSE staff, the "state of the art," as it is presented here, leaves much to be desired. In terms of quantity, of the 429 categories in the three classification schemes, 183 (43%) are empty, and an additional 179 (42%) contain 10 or fewer instruments. In addition, the quality of the instruments, as expressed by their VENTURE evaluations, is prodominately poor to fair....The average ratings for validity, normed excellence, teaching feedback, and retest potential are uniformly poor, while the ratings for examinee appropriateness and usability are predominately fair, with good ratings on these two criteria occuring most frequently in the interpersonal domain and least frequently in the higher-order cognitive domain. In short, much work remains to be done, both in developing instruments where none now exist, and in improving the quality of those instruments which have already been developed. (Hoepfner et. al., 1972, p. 24)

Several other components of the "infrastructure" of tests and measurements should be mentioned (in addition to the Buros Handbooks and CSE Test Evaluations). A number of compilations of measures of classes of variables or specific variables have been published; these have been starred (*) in the bibliography. The Educational Testing Service maintains a library collection of published tests and publishes the Test Collection Bulletin. It should be noted that the needs for instruments for school use and for R & D use are not met equally well. There is a considerable market for standardized tests in the schools, and the "testing industry" makes them readily available, along with scoring services. However, the researcher tends to be concerned with a much broader range of variables than the practitioner, and very often even when an appropriate measure has been developed it has not been published and is not available in quantity.

In addition NIE supports an ERIC Clearinghouse on Tests, Measurement, and Evaluation at ETS which not only provides input to the ERIC system but also commissions "information analysis products". A number of professional organizations give prominent attention to the measurement field, including the American Psychological Association (especially Division 5), the American Educational Research Association (especially Division D), and the National Council on Measurement in Education.

Despite these many services, activities and organizations, it is fair to say that, for one reason or another, many researchers and practitioners still experience great difficulty in locating instruments of specific characteristics for given purposes which have been properly

reviewed and evaluated. This felt need resulted in the formation of the Inter-Association Council on test Reviewing (IACTR) in 1968 (Payne and Watkins, 1973). The Council did a study which did much to identify problems and propose solutions. Unfortunately the organization lacked a firm institutional base and the necessary financial backing and was dissolved in 1972.

Information about the quality of measures is needed by various clients for various purposes. The IACTR experience should be examined carefully to determine whether NIE should sponsor an activity to meet the needs identified by that group. This field might be a prime candidate for establishment of a new institution. None of the laboratories or centers in which NIE supports programs have a major focus on measurement. The closest is the Center for the Study of Evaluation at UCLA, but it deals with evaluation rather than measurement and does not deal with the full range of measures used in educational research and practice. The Buros Handbooks are a personal project of the editor-publisher who is of retirement age and thus lack any institutional base; whether the series will be continued is problematic. In addition there is a good deal of current discussion of the need for item banks and related services. This could be another function of a new institution.

Problems of access to information about tests and measurements exist within NIE as well as in the field generally. An attempt has been made to order the key reference volumes for the NIE library, and the Educational Reference Center provides search and retrieval services. However, especially with a growing intra-mural research program, these

general services may need to be supplemented by somewhat more specialized activities. It is proposed that an information specialist in measurement be added to the staff of the newly formed Educational Reference Division who can assist NIE personnel in obtaining information about tests, research instruments, and specialized collections found elsewhere, such as the ETS collection of published tests, data tapes, items banks, etc.

NIE is beginning the design of a series of periodic and special studies of the R&D system. One element of this program should be the examination of the resources and services available to researchers and practitioners for the improvement of measurement.

Some problems have been noted within NIE in the rigor and consistency with which standards regarding instrumentation have been applied in the review of proposals and the monitoring of projects (Beezer, no date). In the past, some activities have been supported which were not sufficiently rigorous from the measurement perspective. The forms clearance procedure has been concerned primarily with issues of respondent burden and invasion of privacy, not technical adequacy. Proposals focusing on measurement issues have been reviewed very carefully with respect to instrumentation, but often proposals with more substantive foci have been approved even though they provide almost no information on instrumentation. An NIE consultant, John Tuckey, has suggested a system of "circuit riders" who might provide consultative services to principal investigators needing such assistance. This might be helpful, but we also need to introduce more rigor in NIE procedures before proposals are funded. It is proposed that the

agency-wide Task Force on Measurement examine existing procedures for proposal and RFP review in order to strengthen these standards.

To summarize recommendations concerning the availability and quality of measurement instruments:

> RECOMMENDATIONS: In support of its mission to strengthen the scientific and technological foundations of education and build an effective educational research and development system, NIE should support the following infrastructure building activities in the agency and in the field:
>
> - An information specialist in measurement should be added to the staff of the Educational Reference Division to assist staff in obtaining information about tests, research instruments and data sources.
>
> - An instrumentality should be created and supported for expanding and improving the review and evaluation of measurement instruments, including measures of non-cognitive abilities and variables of primary interest to the R&D community.
>
> - An instrumentality should be created and supported which would publish or otherwise make available instruments in the public domain or under license which meet standards of quality and need but for which the market is too thin to invite commercial publication.
>
> - The program of research on the R&D system should include a study of the infrastructure supporting the measurement needs of various agents in the R&D system and make recommendations for meeting other unmet needs through the establishment of new institutions and/or by other means.
>
> - NIE should revise its procedures for review of proposals, RFP's and forms to involve experts on instrumentation and methodology to assure a higher level of technical quality in the research and development supported by NIE in its intramural and extramural programs.

Individual Effects of Testing and Problems of Bias.

The use of tests has grown rapidly since the turn of the century. particularly in the public schools during the last 15 years (Kirkland, 1971). Between 150 million and 250 million tests a year are given, or three to five standardized tests per pupil per year. In additionthere are the external testing programs such as the College Entrance Examination, the National Merit Scholarship, and the American College Testing Program; and the use of tests by industry, business, government, and the military establishment.

Despite this apparent success, testing has increasingly become the object of criticism. These criticisms have eminated from various sectors, including school administrators (Joint Committee on Testing, 1962), and Blacks and other minority groups (R. Williams, 1970). Generally, these criticisms can be divided into three groups: (1) scientific issues concerning the validity of tests; (2) professional issues concerning the misuse of tests; and (3) social issues concerning the consequences of testing.

With reference to validity, Messick and Anderson note that the lower scores typically obtained by minority and disadvantaged individuals may be traced to three possible sources (Messick and Anderson, 1970):

1. The test may measure different things for different groups.
2. The test may involve irrelevant difficulty
   (a) Items that are more germane to one group than to another.

(b) Testing conditions that make some individuals feel anxious, threatened, or alienated.

(c) Differences in test wiseness.

3. The test may accurately reflect ability or achievement levels.

Discussions of cultural bias in tests have emphasized one or another of these factors. Some have gone as far as to propose a moratorium on testing. (R. Williams, 1970). Others have proposed approaches to the elimination of specific sources of bias. There have been various attempts to develop "culture fair" tests of intelligence. Some have translated tests into the primary language of bilingual populations. And others have tried to modify test administration procedures in order to eliminate some kinds of irrelevant difficulty. None of these efforts have been fully satisfactory, and thus a major problem remains with respect to educational programs for bilingual and other sub-cultural groups. There are several programs within NIE for which the problem of bias should be a central issue (e.g. the task forces on bilingual education and the urban disadvantaged). However, it does not seem to be reflected in their plans as yet. NIE should organize a new task force composed of measurement specialists and representatives of relevant R&D programs to plan specific steps to deal with this problem area.

Another set of issues revolves around the misuse of tests. Tests must be used for the purposes for which they were designed and interpreted with reference to the design constraints. Professional standards exist for the development and use of educational and psycholcgical tests (Joint

Committee on Revision of Standards, no date).

One of the problems is that "a test might have a different validity coefficient or a different regression function for a minority/poverty group than for a middle class group and that the general use of prediction equations derived from the White majority might unfairly penalize minority individuals in selection or placement situations," (Messick and Anderson, 1970,).

The National Assessment of Educational Progress (NAEP) has attempted to deal with this problem through the concept of "balancing" (Robert Larson et. al., 1973). National Assessment reports its results in terms of groupings by age, region, sex, size and type of community, color, and level of parental education. Balancing is an adjustment procedure designed to remove the masquerading of one group effect as another and to avoid "double counting" of individuals. An NIE grant is supporting the further development of this method. In a similar vein, Mushkin has proposed the "SIR" (sex, income, race) adjusted index of educational achievement (Mushkin, no date).

Other problems of misuse can be listed (Messick and Anderson, 1970):

- Relevance of the selected test for the proposed purpose
- Side effects (e.g. is test-taking a pleasant or frustrating experience?)
- Misinterpretation of test results (e.g., the presumption that test scores reflect fixed levels of capaiity, or the tendency to take seriously insignificant differences between scores).

- The problem of secondary use, or the use of test results obtained at one point in time for one set of purposes at another point in time for different purposes (raises issues of invasion of privacy, confidentiality of records, and client welfare).

Issues with respect to the effects of testing on students parents, and teachers have been summarized by Kirkland (1971)*:

- Effects on students: What are the effects of tests on the motivation, self-esteem, and self-perceptions of students? Do they affect study habits and teacher-pupil relationships? Do they produce anxiety and emotional tensions? Are pressures to achieve by teachers, parents, and schools made as a result of tests? Do tests encourage dishonesty in the form of cheating, faking, etc.? Do they create labels of inferior or superior intellectual status? Do they determine one's adult social status? What advice is given students on the part of parents, teachers, and schools as a result of test scores? What is the influence of tests on the opportunities open to individuals?

- Effects on parents: What are the effects of tests on parents? Do their children's test experiences produce tension and anxiety in them? Does the importance that tests have in selection and placement cause parents to inflict undue pressures on their children? Does knowledge gained from their scores influence parents' perceptions of their children's abilities? Does this knowledge influence the advice parents give to their children?

- Effects on teachers: Are pressures placed upon teachers as a result of tests? Do tests determine teaching and evaluation methods? Are teachers evaluated by these tests? Do they color the teachers' perceptions of students? As a result of tests, do teachers behave differently toward students?

In many respects the questions about the effects of testing on the life chances of individuals are among the most serious raised. It is charged that tests may predict the ability to do well in school, but

*Systems effects of testing are discussed in a later section.

neither test results nor school grades predict success in occupations (Jencks, 1972, Berg, 1970). This is as much a criticism of schooling as it is of testing, although industrial use of testing is not notably more successful. Of course it is not the sole function of the schools to prepare people for jobs. But we need to be able to define the knowledge and skills of a competent adult in his various roles and to be able to determine whether the schools are making their proper contribution toward education for adulthood. (Mobility issues are discussed further in the section on measurement of systems.)

Granted that there are a number of problems associated with the use of testing, there would also be social consequences of not testing, as Messick and Anderson point out (Messick and Anderson, 1970). The risk is that subjective forms of appraisal would be substituted with the likelihood that bias and discrimination would increase.

> The elimination of tests would also mean the loss of one of the best ways for teachers to acquire a useful appreciation of the broad range of competencies and traits that characterize human behavior or to develop needed sensitivities to the nuances of cognitive growth.* An increased parochialism might spread throughout education because of the absence of a national normative perspective and the limination of access to concrete examples of what other educators deem important to assess. And of utmost importance,

*A reviewer of an earlier draft of this paper takes issue with this point, feeling that the use of test tends to narrow the sensitivities of teachers. The difference may be between the potential use of tests and what happens more typically. This issue would be worth investigation empirically.

> there would be an absence of yardsticks for gauging the effectiveness of educational programs and for evaluating the equity of the educational system. (Messick and Anderson, 1970, p. 87).

With reference to the entire range of problems identified under the heading of "effects of tests and the problem of bias," a number of current activities are worth noting.

- A revision of the "Standards for Development and Use of Educational and Psychological Tests" is now in the fourth draft of a revision under the sponsorship of three professional associations.*

- In the spring of 1973 a National Workshop on Testing in Education and Employment was organized to focus on the need for reform in procedures for testing racial, ethnic, and low socioeconomic groups in America.

- NIE made a grant in June 1973 to support, in cooperation with three foundations, a project designed to study the effects of testing in Ireland. Hitherto Ireland has not used standardized tests in its schools. A decision has now been made to introduce testing, and the project represents an agreed-upon plan to do so under an experimental design. The two main foci of the research are (1) to study the consequences of introducing testing at individual institutional, and cultural levels, and (2) to do a case study of the research as an instance of planned social experimentation.

- NIE has a legislative mandate to build an effective research and development system, and the Planning and Policy Analysis Unit of the Office of Research and Development Resources is undertaking policy studies to determine how best to fulfill that mandate. Testing and the testing industry are part of that system and will be included in a survey of the R & D system now being designed.

---

*The Joint Committee on Revision of Standards includes representation from the APA Committee on Psychological Tests, the APA Board of Scientific Affairs Liaison, the American Educational Research Association, and the National Council on Measurement in Education.

RECOMMENDATION:

- An Exploratory Studies group, working through the agency-wide Task Force on Measurement should give high priority to a research program on the effects of testing and the problem of bias, working with the Task Force on Bilingual Education, the Task Force on Education for the Urban Disadvantaged, and other relevant units.

Theoretical and Methodological Issues

Any detailed treatment of theoretical and methodological issues tends to become quite technical and is probably beyond the competence of the present author. There have been a number of recent statements summarizing the state-of-the-art and pinpointing areas where new work is needed (Cronbach, 1970; Thorndike, 1971; McClelland, 1973; Ebel, 1973; Kirkland, 1971; Krantz, et. al., 1972; Anderson, Messick, and Hartshorne, 1972). Certainly the field is in ferment, both in education specifically and the underlying behavioral sciences generally. Here we will attempt only a brief listing of some of the salient problem areas.

In the last ten years the concepts of criterion referenced testing (Glaser, 1963; Popham and Husek, 1969), domain-referenced measures (Hively, et. al., 1973), and mastery learning (Block, 1973) have emerged. The literature on these topics is still rather confused, and their value for the improvement of education and educational research has yet to be determined. Nevertheless, that potential is sufficiently challenging to warrant NIE support of continued work in these fields.

The use of standardized tests developed to measure individual differences for the purpose of evaluating educational programs has become a controversial area. Fennessey has reviewed these issues and suggests

that their use may be quite appropriate under certain conditions (Fennessey, 1973).

There has been an increasing dissatisfaction with cross-sectional research designs and a growing interest in longitudinal research. This requires the development of new methodologies appropriate to the measurement of change. Several activities in this area should be noted. Two federal interagency committees, one on early childhood and one on adolescence, have supported a special interest group on longitudinal research. The group has identified some of the problems of longitudinal/intervention research and compiled information on important studies now underway. (Grotberg and Searcy, 1972; Grotberg, 1972; Lazar, 1972). They are now holding discussions concerning the possible use of "marker variables," i.e., agreed upon measures of key variables which would be used in all related projects (regardless of what other measures were used) so as to provide a link between similar studies and promote the cumulation of knowledge. Trent and his associates have also compiled information about longitudinal studies and done an analytic comparison of their conceptual frameworks, methodologies and findings. (Trent et. al., 1972-73, 5 Vols.) Finally, the Board on Human Resources of the National Research Council has been examining and comparing various data sets available from projects conducting longitudinal studies and from professional associations which do studies of their membership.

This section on measurement of individuals perhaps has focused too much on the use of tests and the concerns of psychometricians. There are other methods of collecting information, although some may be more

relevant to the research worker than to the practitioner. Observational methods both in the field and the laboratory are important tools of data collection, and many categorical schemes have been devised for classifying behavior and interactions (Simon and Boyer, 1967 and 1970). The field notes and participant observation of the anthropologist and the sociologist need to be considered. Despite various problems and criticisms, the survey is still a widely used research tool. The recently established Social Science Research Council Center for Social Indicators is now attempting to achieve consensus on the wording of a set of "background variables" such as education, occupation and marital status in order to improve the comparability of data among surveys. The logic and methods of survey analysis have been improved and refined over the past twenty years. The interview provides great richness of detail and depth of meaning, perhaps at the expense of comparability, but at certain stages of research such data can be the source of great insight. The imaginative use of school records, financial accounts, and administrative statistics can provide valuable information. Such data fall in an important class of unobtrusive measures (Webb, Campbell, Schwartz and Sechrest, 1966) which have the methodological advantage of being nonreactive, i.e., they do not tend to modify the behavior of the person being studied. On the other hand there are problems for which physiological measures may be quite appropriate.

There is a considerable literature about each of these methods, and a separate paper could be written about the advantages and problems of each. For the moment we will have to content ourselves with the admonition

to NIE program managers to choose the method to fit the problem and to recruit staff members from an appropriate range of disciplines and methodological traditions.

In FY 73 NIE supported a program then called Field Initiated Studies in which one of the panels focused on "objectives, measurement, and evaluation". The total program provided $10,285,000 (commitments for FY 73) in support of 193 projects. Of this, $917,492 went to 29 projects concerned with objectives, measurement and evaluation.

Under FY 74 plans for support of field initiated research, to be administered by the Office of Research Grants, consideration is being given to dissolving the Panel on Objectives, Measurement and Evaluation and instead assigning proposals in the this field to other panels. From the perspective of this paper such a step would be unfortunate. A separate program area on the theory and methodology for educational measurement is needed because panels reviewing proposals on substantive problems concentrate on those problems as such. They tend to be satisfied with current methods, even methods with known limitations, rather than insist on confronting and resolving methodological difficulties. Further, the support of field initiated research should be considered a key strategy for support of theoretical and methodological problems. It is a relatively non-mission oriented aspect of educational R & D and one in which maximum freedom for the investigator is generally viewed as being most productive.

RECOMMENDATIONS:

- The Office of Research Grants should maintain the identity of the Panel on Objectives, Measurement, and Evaluation and should provide leadership through conferences and other activities in making NIE's interest in this field known to the research community and otherwise stimulating a larger number of high quality proposals.

- The Task Force on Measurement should undertake or support relevant instrumentation studies when the need for specific research is identified in connection with mission-oriented programs. This should include research on instrumentation problems associated with longitudinal research and the measurement of change.

Measurement of Systems

Educational systems are also important units of analysis in the study and practice of education. The measurement of individuals is the province of psychometrics and is relatively well developed. Concern for the functioning of systems is a more recent phenomenon and the development of theory, identification of the relevant variables, and the formulation of appropriate measures is much less advanced. Here measurement specialists and theorists in sociology, political science, economics, education and anthropology have much to contribute.

One may be concerned with the functioning of educational systems at any of several levels. Although the classroom level is often thought of as the lowest level of analysis, there are smaller units of some importance: the peer group, teams of professionals and paraprofessionals; pupil teacher dyads, or other units. Above the classroom are the school, the school district, state, and nation, with intermediate levels sometimes of interest. The existence of multiple levels means that the status of a given variable may change from problem to problem. For example, what is a dependent variable in one problem may be a contextual variable in another.

Programs and Processes

For reasons difficult to divine, little is known about what goes on in America's classrooms. Perhaps the nature of major federal programs

has had something to do with this. Most have been based on a model under which resources were supplied to support unnamed and undescribed "innovations" to alleviate some educational problem. The innovation to be attempted was left to local choice and initiative. While guidelines were provided and there were criteria for rejecting projects, the actual nature of the treatments supported covered a very considerable range. Often innovations have existed largely at the label level, with no common understanding of what the specifications for the innovation were. Thus terms like "teacher centers", "open education", "team teaching", "educational renewal", and "differentiated staffing" have been little more than conceptual inkblots to some, with each teacher, school, school district or federal official supplying his own meaning to these terms.

Educational developers, such as those in the laboratories, have had to be more concerned with the nature of their treatments, for that is what they were inventing. However, where their new products were tested in comparison with "traditional" practice the attempt to describe "traditional practice" in detail and its similarities and differences with the new product has often been lacking.

At the national level there is little known about the nature of educational practice. In looking at the literature one sometimes gets the impression that the schools are the same as they were 30 or 50 years ago, while at others it would appear that very substantial changes have taken place in a majority of schools. Perhaps both statements have some validity, but they refer to different aspects of practice. What are the

facts? A specific project that NIE might wish to consider would be a national sample survey of education practices. The feasibility of such a study would depend largely on the ability to develop suitable measures of educational programs and processes.

What might be the major facets of such a study? In describing what he calls the "means of education". Bruce Joyce differentiates three systems: (Joyce, 1969).

A. The social system of the school
   1. The normative structure
   2. Student roles
   3. Teacher roles

B. The technical support systems
   1. Data storage and retrieval systems
   2. Instructional systems
   3. Information processing systems
   4. Materials creation and consultation systems

C. Curriculum systems
   1. Content of subjects or curriculum areas
   2. Sequence
   3. Repetition of ideas, principles or values to provide continuity
   4. Teaching strategies
   5. Mode of presentation
   6. Assessment and feedback systems

Of course these systems and components interact with one another. The phenomena represent very different measurement problems, and the existence of appropriate measures is quite uneven. Price has assembled a compendium of operational measures of organizations (Price, 1973). Dreeben has opened up some of the conceptual and theoretical problems of the normative outcomes of schooling (Dreeben, 1968a and 1968b). He places schooling in the context of socialization and argues that some of the important norms

learned in school are a function of the social system of the school rather than the explicit curriculum. Some specific measures of the normative structure are covered by Mason (1953). Many aspects of educational practice can best be measured by observational methods. Research for Better Schools has assembled information about a large number of interaction analysis schemes (Simon and Boyer, 1967 and 1970). Corwin has developed Guttman and Likert scales and other indices to measure structural and group characteristics of schools, including standardization, centralization of decision-making, patterns of supervision, group cohesiveness, and professional and employee role conceptions of teachers. (Corwin, 1970). Boocock and Cohen have each contributed to the conceptualization of sociological variables at the school and classroom level as related to student learning (Boocock 1966 and 1973; Cohen, 1972).

Educational researchers and policy makers have tended to confine their attention to the formal school system. Within that framework, in terms of programs and processes there is more known about what goes on in elementary and secondary schools than about post secondary education. But outside of the formal school system there are tremendous amounts of educational activity that take place in other settings: in employer operated programs (e.g., NIE's Career Education Model II), in the armed services, in the home through correspondence, television, and open university programs, in evening schools, proprietary schools, etc. etc. Just as we must begin to understand schooling in relation to the total socialization process (i.e. all the processes which determine how the

young become adults), we must be able to place "establishment schooling" in relation to other explicitly educational institutions in our society.

The evaluation of educational programs has become a very important type of inquiry in education. Some studies reflect a school of thought that focuses almost exclusively on outcomes. When many studies have indicated program failure or partial failure, the question of "why?" inevitably arises. To make such causal inferences requires a different kind of evaluation design, often referred to as evaluation research (Suchman, 1967; Rossi and Williams, 1972). Such work requires the conceptualization of different classes of variables and their interrelationships. The identification of processes and programs becomes crucial in such designs.

A frequent finding in evaluation studies is that the innovation or product was not actually implemented in the manner specified by the developer (Gross, 1971; Solomon et al, 1973). Clearly it is not enough to use the developer's specifications as the measures of program and process; the researcher must get into the classroom and determine what is actually happening.

According to Suchman there are two possible sources of program failure.

> If a program is unsuccessful, it may be because the program failed to 'operationalize' the theory, or because the theory itself was deficient. One may be highly successful in putting a program into operation but, if the theory is incorrect or not adequately translated into action, the desired changes may not be forthcoming: i.e., "the operation was a success but the patient died." Furthermore, in very few cases do action

> or service programs directly attack the ultimate objective. Rather they attempt to change the intermediate process which is 'causally' related to the ultimate objective. Thus, there are two possible sources of failure (1) the inability of the program to influence the 'causall variable, or (2) the invalidity of the theory linking the 'causal' variable to the desired objective. We may diagram these two types of failure as follows:
>
> | INDEPENDENT VARIABLE | INTERVENING VARIABLE | DEPENDENT VARIABLE |
> |---|---|---|
> | Activity or Program | 'Causal' Process | Desired Effect |
>
> Program Failure — Theory Failure
>
> According to this analysis, evaluative research tests the ability of a program to affect the intervening 'causal' process. Non-evaluative or basic research, in turn, tests the validity of the intervening 'causal' process as a determinant of the desired effect. (Suchman, 1971)

Some investigators hold that it is at the point of interaction between aptitude or trait and the treatment that great promise lies for improving our understanding of the educational process (Cronbach, 1970). The general notion is that there is no one "best" instructional program for all students; rather, characteristics of students (e.g. personality, ability or status variables) can be identified which exhibit differential relationships with characteristics of treatments (e.g. inductive vs. deductive or structured vs. unstructured). While a number of such interactions have been found, most have not yet been replicated, and there are many cases where hypotheses about interactions were not confirmed (Berliner and Cahen, 1973).

Perhaps the person variables have been studied more carefully than the treatment variables. Such research cannot succeed unless the conceptualization and measurement of the treatments is equally sophisticated and rigorous. We need to identify the important dimensions that characterize educational treatments, develop a methodology for quantifying them, and determine their usefulness for comparative evaluation.

For one thing, treatments cannot be reduced to curriculum materials and teacher behavior; the social organization of the school and classroom must also be understood. Concepts such as peer group, school climate, role structure, compliance and control mechanisms, and type of grouping are among those of importance. There is increasing experimentation with the organizational aspects of education as witnessed by innovations in team teaching, differentiated staffing, and open education. However, much more needs to be known about the relation between group structure and process, on the one hand, and social psychological concequences in behavior on the other. Such research will require measures of qualitative relationship's within the learning group, over time.

NIE has sponsored work on the multi-unit school at the University of Wisconsin Research and Development Center for Cognitive Learning, as well as work on a variety of organization effects at the Center for Social Organization of Schools at Johns Hopkins University. Some of the research in this field has been reviewed by Boocock and Cohen (Boocock, 1966 and 1973; Cohen, 1972).

RECOMMENDATIONS:

- NIE should design and conduct a national survey of educational practice which would determine what educational materials, methods, organizations, and technologies are actually in use, identify innovative or experimental programs, and determine the number of pupils and instructional staff involved with different practices.

- NIE should inauguarate a program of research and policy studies designed to describe and understand educational institutions and programs which fall outside the formal school system should identify and study salient policy issues concerning the relationship between the formal and informal systems.

- NIE should use evaluation designs which provide careful measurement of treatments and the degree of their implementation and should support the development of such measures where appropriate. Explanatory models of evaluation are to be preferred.

- NIE should give some priority in intramural and extramural research programs to two substantive areas: (a) aptitute-treatment interaction (ATI) or trait-treatment interaction (TTI) studies, and studies of the sociology of learning, i.e. studies of the effects of social and organizational factors on learning.

Inputs and Contexts

The economic, political, ethnic, racial, community, cultural and social systems in which schools and colleges are embedded provide important inputs and contexts for the understanding of education.

Despite the apparent finding that variations in economic resources have little effect on educational outcomes (Jencks, 1972, Coleman, 1966, Spady, 1973), it is difficult to believe that there will not be a continuing effort to study the effect of resource allocations. The key to this would seem to be to move away from the conception of the school

as a black box into which resources are fed at one end, and out of which educational results issue from the other; the question is what is the money spent for, and what are the efficiencies of various uses of resources? More recently Coleman has advanced a theoretical framework for studying social change in terms of the conversion of resources (Coleman, 1971). Certainly the allocation of economic resources will continue to be an important political and social issue as a matter of equity quite apart from research results or lack of results.

The Coleman Report and other studies have also pointed to the rather large and stable effects attributable to family and community factors, particularly socio-economic status. What is not so generally recognized is that the variables used in such studies are mostly proxy variables; it is difficult to infer directly from father's occupation (for example) to achievement tests results. Through what processes and intervening variables are such effects produced? It is in this area that understanding must be achieved before interventions can be designed. There are substantial bodies of basic research literature which can be focused on this problem which center on concepts and processes such as socialization (Goslin, 1969, Inkeles, 1966 and 1969, Coleman, 1972); Self-concept/self-esteem (Crandall, 1973,Langenfeld, 1972); Social competency (Anderson and Messick, 1973, McClelland, 1973); Social stratification (Duncan, 1968).

The importance of the larger sociocultural environment in influencing formal education and the outcomes of schooling is the major theoretical orientation of an important new book by Herriott and Hodgkins.

> Given such a perspective, the conclusions of many studies that "educational outcomes" are more likely a function of factors outside of the school than of those within it, take on a new meaning, for they illustrate the more general fact that as "open" social systems, educational organizations are continually influenced by society. Thus, it is not simply that children within the educational system fail to learn, but rather that what they learn is determined in large measure by the interaction of school and society. (Italics in original) (Herriott and Hodgkins, 1973, p. 15)

Two recommendations for NIE relate to inputs and contexts:

> RECOMMENDATION: NIE should support the development and standardization of input and context variables as a means of achieving greater understanding of the effects of these factors on educational experience and as an aid in improving the comparability and cumulativeness of educational research. In addition, NIE should develop a research agenda focused on the influence of elemants on the larger society on formal education. (See also the discussion of monitoring indicators of social change below).

Outputs and Indicators

There are several current and salient strands of thought that have focused attention on the need for systems level output measures. Within education there has been a call for greater accountability in the various sectors of the enterprise (Stake, 1973; Levin, 1962). Among social intervention programs generally the need felt for program evaluation has stimulated considerable intellectual ferment and a whole new "evaluation industry" (Wholey et al., 1970; Rossi and Williams, 1972; Suchman, 1967). And a concern for understanding the meaning of rapid social change and

planning for the future have been principle factors behind the work done on social indicators (Sheldon and Moore, 1968; HEW, 1969; Land, 1972). All three lines of inquiry share a focus on the need for systematic data basic to social policy decisions.

While there is considerable overlap in the domains encompassed by each of these concepts, each has a somewhat distinctive perspective or emphasis. The work on accountability tends to fall within the management framework of the operating school system. What data do we have to measure the effectiveness of our schools and school systems? (This concept also reaches down to the individual level in its concern for accountability of administrators and teachers). Program evaluation tends to take on the perspective of the Federal, state, or foundation program manager who is administering "categorical" funds. Such programs cut across operating organizations, introducing some incremental innovation in each. Those who have used the concept of social indicators have tended to be concerned with the operation of our institutions at the most macroscopic level. Thus somewhat different conceptual frameworks have evolved around the need to systematize policy decisions at each level of the system.

The development of organizational output measures is still fairly primative, both conceptually and methodologically. The tendency is for each investigator to develop his own measures, and often little work is done to determine their validity or reliability. The listing of compilations of measures in the bibliography include systems level measures as

well as individual level measures. Also many of the basic facts of educational attainment, degrees, etc., are collected on a systematic and comparable basis by the National Center for Educational Statistics of OE and by the Census Bureau, both in its decennial census and the monthly currently population survey. However, there is little agreement among researchers on the measurement of direct systems variables of analytic interest.

There are a number of activities of current interest in and outside of NIE dealing with outputs and indicators at the systems level.

The National Center for Higher Education Management Systems has been developing management information systems for use by colleges and universities. To date the work has included largely cost and other administrative data, but they are moving more toward the measurement of the benefits required by cost/benefit analysis. Some of the work on outputs of higher education is covered in Lawrence et. al., 1970.

Abt Associates, the evaluation contractor for the rural schools within the Experimental Schools Program, is using a sophisticated conceptualization of organizational change (based on general systems theory-organizational environment, input, throughput, output, structure and culture-and change stages-evaluation, initiation, implementation and routinization) and has identified appropriate measures for its components.

The National Assessment of Education Progress (NAEP) is an ambitious attempt to ascertain the knowledge, skills, understandings and attitudes

of young Americans (Womer, 1970). Four age levels are sampled: (9-year-olds, 13-year-olds, 17-year-olds, and young adults between the ages of 26 and 35) and ten subject areas. The focus is on the measurement of attainment in an absolute sense rather than with reference to some norm: what proportion of a given group possesses a given skill or knowledge? The sampling is done in a manner which does not permit reporting of results by school, school district, or state; rather results are reported by region, size and type of community, sex, color, and parental education. This limitation seemed to be necessary in order to establish the program because of the sensitivities of states and school districts. However, a number of states have now used the model to implement state assessment systems (including Michigan, Maine, and Pennsylvania).

An important objective of the program is to measure change over time. In any one year data are collected on only two subject areas, and each area is reassessed approximately every five. Some of the items are repeated in each cycle, and so it is possible to determine whether the level of attainment of an age group is increasing over time. The second cycle of data gathering has begun for some subject areas, and change data will soon be available.

These data are useful for a variety of purposes other than descriptive monitoring, including analyses of curriculum content. The NAEP staff publishes many helpful reports but does not claim to be exploiting the complete potential of the data. This is one of a number of data sets which various programs in NIE could make valuable use of for

secondary analysis purposes. Some of the methodological problems of secondary analysis of sample surveys are covered by Hyman, (1972).

There is an important limitation to the value of the NAEP data, namely the paucity of data on other variables to use in its analysis and interpretation. Some limited information is collected on background variables (e.g. age, sex, region), but no information on the nature of the educational programs to which respondents have been exposed. Thus NAEP must be classified an another example of "black box" research which fails to include important educational variables. It is granted that there may be difficulties in collecting such information, given the constraints under which the project operates. Possibly such analyses can be performed on data collected in some states and local school districts which have patterned their assessment systems after NAEP.

Representative Albert Quie has introduced legislation which would make the methodology of National Assessment the basis for a major change in the manner of distributing Federal funds for the disadvantaged (HR 5163). Until now the funds for Title I of the Elementary and Secondary Education Act have been distributed on the basis of economic indicators, used as proxies for educational disadvantagement. Quie notes the lack of a perfect correlation between economic and educational measures of disadvantagement and proposes that the distribution should be based on direct educational measures. His bill would require collecting NAEP type data on reading and mathematics on a basis which would permit reporting of results for each state. The individual states would, in turn, be

required to implement state assessment programs which would be the basis for allocating the funds to local districts. NIE has been discussing the implications of this approach with some of the experts in the field. While there are many attractive features to the proposal, are a number of problem areas in which research would be highly desirable before becoming committed to a large scale national program having great significance in the allocation of large amounts of federal support. Among the more important are the following (Madaus and Elmore, 1973):

- What would be the effects of the negative incentive feature of the bill and how would they compare with a positive incentive system?
- What is known about the effects of other external testing programs, particularly the problem of "teaching to the test" and whether such programs have the effect of inhibiting innovation and homogenizing the educational program?
- In the several states that have adopted state assessment systems, what has been the effect of such systems, especially in Michigan where the system is used to allocate resources?
- How can NAEP type data be aggregated and summarized, and what are the methodological problems involved in setting performance standards?

The social indicators movement has had a short and checkered histroy (Brooks, 1972). While there is considerable disagreement on the meaning of the term, there seems to be consensus on several elements: (a) social indicators are time series data which permit the monitoring of change over extended periods of time permitting the separation of long term trends from short term fluctuations; (b) they may be either quantitative or qualitative; and (c) they can be disaggregated by relevant attributes of either the

persons or the conditions measured (such as skin color or year of construction) and by the contextual characteristics that surround the measure (such as region or city size) (Sheldon and Freeman, 1970). Among the early hopes were that a system of social accounts could be developed comparable to the system of economic accounts, and that the indicators would be directly useful for program evaluation and the setting of goals and priorities (National Commission on Technology, Automation and Economic Progress, 1966; HEW, 1969). Senator Mondale has introduced legislation which would establish a Council of Social Advisors responsible for preparing an Annual Social Report to the President.

More recently some of these early statements of expectations have been criticized as unsound and unrealistic (Sheldon and Freeman, 1970; Sheldon and Land, 1972). The social area lacks a common metric and a model of the social system from which to derive a system of social accounts. Social indicators are the product of multiple causes, and the effects of specific government programs cannot be disentangled from other causes. The setting of goals and priorities ultimately depend on value choices not the assembling of data.

Nevertheless there seems to be agreement that the concept is still useful in relation to the key function of monitoring social change, both in its objective dimensions (Sheldon and Moore, 1968) and subjective dimensions (Campbell and Converse, 1972). It is also helpful in pointing to the need for standardization of measures in the social field. Furthermore, we are beginning to see the development of models of systems

or sub-systems which provide some understanding of causal networks (Land, 1972; Anderson, 1973). The full potential of the social indicators concept will not be reached until the indicators can be integrated into explanatory models and theories; but this advance in turn may be dependent on the development and improvement of appropriate measures.

Currently the Office of Management and Budget is circulating a draft social indicators report which would be issued on a periodic basis. The education section of the report consists of Census and OE data on enrollment, retention, graduates, and degrees plus some of the National Assessment results.

The National Science Foundation sponsors a program of research on social indicators. Three projects or activities of special interest to NIE are: (1) development of a framework for national goals accounting (National Planning Association, 1972); (2) support for the Social Science Research Council's Center for Coordination of Research on Social Indicators (which among other things is seeking to standardize the wording of a number of "face sheet" items frequently used in sample surveys); and (3) several projects to develop uniform measures of social competence.

The importance of non-educational indicators for NIE lies in the fact that many of the major changes in education have come about in reaction to forces originating outside of the educational sector, such as Supreme Court decisions, the "baby boom," Sputnik, the war on poverty, the movement for community control, concern with youth unemployment, and the

movement of women into the labor force (C. Williams, 1973). Also, in a rapidly changing society the schools of today need to anticipate the nature of the society which tomorrow's graduates will be entering. Programs are needed within NIE which focus on the interface between education and other key sectors and seek to prepare students for tomorrow's world.

Educational indicators are, of course, a type of social indicator. The educational field is relatively rich in the number of statistical time series available through the National Center for Education Statistics (NCES) of the Office of Education, the Census Bureau and various state and local educational agencies. However, the indicators available vary considerably in their usefullness for either theory or practice. For example, while there is considerable information about inputs and of gross outputs like graduates and educational attainment, there is little specific information about educational practice or on the knowledge, attitudes and behaviors of pupils and students.

Granted this problem, there are more time series available than are being properly exploited. Important data are often available at state and local levels when not available nationally. Fortunately the situation is beginning to change, and a few efforts can be cited which show how such data might be used and how they need to be improved. Abbott L. Ferriss has been one of the leading workers in this particular vineyard. In 1969 he published Indicators of Trends in American Education in which he reviewed a large number of the time series and identified significant

trends that were observable. He has urged the usefulness of such data in serving a monitorship function (Ferriss, 1972). Monitoring consists of determing whether new observations represent the continuation of past trends or whether they signal a turning point. If the latter, the task is to determine whether the change has significant consequences for the future, particularly for other normatively significant elements in the system. Clearly this kind of function is essential to any policy analysis activity in NIE.

Ferriss has suggested that there are at least four types of educational indicators that would be highly useful for monitorship, providing clues to intervention:

- Measures of the educational status of the population, primarily the out-of-school population; for example, ideally this would be an inventory of the skills in the population; practically as a minimum we now can determine the following: years of school completed (by various traits, such as age, sex, color, etc.), percent of the population with various degrees, by field, percent of the population certificated at given levels of competence by various professions, etc.

- Educational progress of the school population: continuation ratios by age, sex, color, etc.; grade progression; dropout rates; completion rates, etc.

- Qualitative information on the staff of educational institutions.

- Measures of characteristics of the school. Characteristics chosen should possess demonstrated relationships to educational outcomes, that is, that are dicated by explanatory models and theories. (Ferriss, personal communication)

NCES has been concerned with rationalizing its statistical system and has commissioned a number of papers by Selma Mushkin of Georgetown University to explore the problem of output measures in education (Mushkin 1971; 1972a; 1973). A recent product has been Indicators of Educational Outcome Fall 1972 (Cobern, Salem and Mushkin, 1973), which includes a classification of outputs of potential value (see Table).

Table A. -- Summary Classification of Outputs* With Selected Examples

Time Phase 1 (Primary Effects)

Product Consumption		Investment	
Quantity	Quality	Income	Employment
Number of students, High school completions, etc.	Attitudes, Attributes, Aptitudes, Achievements (e.g., self-esteem, creativity, IQ, SAT scores)	Value added, Earnings, Added earnings, etc.	School dropouts, Unemployment rate, etc.

Time Phase 2 (Secondary Effects)

Investment Feedback	Consumption Feedback
Economic growth (e.g. Years of schooling, lifetime earnings differentials)	Consumer information, Consumer efficiency, Medical care use, Use of leisure time, Moral and citizenship values, etc.

Time Phase 3 (Tertiary Effects)

Intergenerational Impacts

Educational motivation of children

* In addition to benefits to students, there are benefits to parents such as the babysitting or child care activities of the school.

Source: Cobern, Salem and Mushkin, 1973, p. 7.

NCES is also the sponsor of the National Assessment of Educational Progress, discussed earlier, which will provide useful education indicators once the cycle of data gathering starts to produce time series results. Some agreed-upon way of computing summary scores is also needed.

The Office of Education has been sponsoring a program of monitoring social trends at the Educational Policy Research Center at the Stanford Research Institute (SRI). This work is based on a "future research" framework (William, C. 1973). NIE needs to develop some formal ties to

to this program.

RECOMMENDATIONS: NIE should organize a small staff in the Planning and Policy Analysis Unit to monitor social and educational change through such activities as:

- Analysis of educational and social indicators published by other agencies
- Conduct and support for projects building explanatory models of the educational system and the larger social system in which it is embedded
- Identification and refinement of measures of variables needed in the models
- Liaison with organizations collecting indicator data, with the OMB social indicator unit, the SRI Center, and other relevant organizations
- Support for special extramural studies of the impact of outside forces on education
- Serve as an information resource for the National Council on Educational Research

All groups in NIE need to be as sensitive to the need for systems measures as to individual measures for the understanding of programs, processes, inputs, contexts, outputs and indicators. Recommendations made in the previous section of the paper regarding the activities of the agency-wide Task Force on Measurement and the Exploratory Studies Group on Measurement, Methodology and Secondary Analysis should be expanded to encompass the need to improve our measurement of systems.

Systems Effects of Testing

It is not easy to separate the issues surrounding the effects of testing into individual and system effects since many individual effects have system consequences when aggregated. For this reason many of the points made in the earlier section on individual effects of testing and the

problem of bias are relevant here as well. Nevertheless it will be useful to refocus our attention on the problem from the systems perspective.

One of the features of social change in the past ten years has been the decline of the melting pot philosophy and the growth of cultural pluralism. Some years ago Florence Kluckhohn pointed out that not all departures from dominant culture patterns are deviant, i.e. "bad" (Kluckhohn, 1953). Any society, particularly one as complex and heterogeneous as ours legitimizes departures from the most common modes of behavior for certain groups and roles under certain circumstances. Thus we have both dominant and variant culture patterns which are viewed as legitimate. We have been witnessing the proliferation of variant culture patterns in the United States during the past decade.

Problems arise when the construction, use, or interpretation of tests or other measures is not anchored in an appropriate cultural frame of reference (ETS, 1973). Standardized tests which have been normed on white middle class populations might be quite invalid if used to assess the general ability of a lower class black; yet if we shift the frame of reference it might be quite accurate in reflecting the assimilation of the lower class black into the dominant culture. By the same token, the "BITCH Test" (Black Intelligence Test of Cultural Homogeneity-Education Daily) may be an accurate reflection of intelligence of those raised within a particular ghetto sub-culture, it would be useless for either blacks or whites in relation to any activities outside of the sub-culture.

So what do we mean by cultural pluralism or variant cultures? Some advocates of "bilingual-bicultural" education speak as if we have or would like to achieve multiple parallel societies such as those found in Quebec or Belgium. Certainly this is implied when they advocate school programs in which a full curriculum in Spanish is offered K-12 in parallel with an English curriculum for all pupils. However, such a parity is not now reflected in occupational and other spheres of our society and is not likely to be in the forseeable future. Indeed, one suspects that the chief goal of most minority group parents, whatever their pride in their own ethnic heritage, is for their children to become full members of the majority society, at least in their occupational roles. The point is to recognize that there is no inconsistency between the parallel existence of dominant and variant cultures so long as one can sort out which is appropriate in various times and circumstances.

Questions have been raised concerning the use of tests and other assessment procedures to serve gatekeeping functions in the stratification system: sorting children into different tracks or curricula within the school system; selecting those to be admitted to college; and selecting those for admission to or placement within the occupational world. Some critics feel that the system is too decisive at various points and argue for keeping options open for longer periods. Furthermore, the selection process is difficult to defend when evidence is often lacking that the criteria used to sort and select have direct relevance to later occupational success and may often mislabel young people on the

probability of educational success.

While the arguments against educational selection often seem compelling, we need to proceed with caution.

> ... we should not overlook two possibilities: that our schools and colleges generally may be more meritocratic--use more universal standards for advancement--than the world of work; and that loosening the meritocratic or allocative function of education may create more inequality of opportunity than presently exists, leaving the most important educational decisions (e.g., who goes to college and where) to fall once again upon the family, social heredity, or politics. If indeed our economic system arbitrarily discriminates against racial, sex, and other "minorities" to the extent that some observers have indicated, one could argue for more rather than less universalistic standards in educational selection and a closer rather than a looser fit between educational attainment and occupational placement. At least we should proceed cautiously in condemning our schools and colleges for setting standards which not everyone is expected to achieve. Unlike the world of work where the norms of achievement are frequently and perhaps necessarily evaded (e.g., in job rights and seniority), schools may be the more important arena for "letting the best man win." (Clark, 1971)

We have already alluded to the possible effects of external testing programs in discouraging innovation or departure from a dominant core of content. The National Assessment of Educational Progress, it should be noted, employs an elaborate process of identifying consensus objectives on which to base their exercises. It would be important to determine whether such a methodology has a rigidifying effect on school programs, either in connection with NAEP itself, or the use of comparable assessment

systems at state or local levels.

RECOMMENDATION:

An Exploratory Studies group should undertake a program of research on the effects of testing and other assessment methods which would study such problems as:

. How does the selection and channeling process now operate in schools and how can it be improved? What is its effect on different cultural sub-groups in the popultion? Do tests foster a narrow conception of ability and reduce the diversity of talent available to schools and society?

. What effect does testing have on the diversity and innovativeness of school programs? Do new technologies like the use of item banks and does computer testing provide solutions to problems posed by older methods?

This research program should not be conducted in isolation from other NIE activities, but rather should work through the agency-wide Task Force on Measurement and "piggyback" on other programs, such as those dealing with bilingual education, education for the urban disadvantaged, and the evaluation of experimental schools, wherever possible and appropriate. Some of these issues will be studied using a unique experimental design in the Boston College project examining the introducting of testing in Ireland.

Theoretical and Methodological Issues

As with measurement of individuals, we will eschew a detailed treatment of theoretical and methodological issues concerning the measurement of systems. Instead we will content ourselves with noting

some of the different types of measures encountered at the systems level and citing a number of recent papers which discuss some of the principal methodological issues.

A rough categorization of types of measures would include (a) aggregated data, or characteristics measured by summing data from individuals or lower order systems; (b) context data, or data characterizing higher order systems; (c) direct systems measures, or characteristics which are not derivative of either lower or higher order systems; and (d) derived measures, or measures such as ratios which represent relationships between other variables. Frequently it happens that the investigator concerned with one level of analysis is forced to adjust data obtained at a different level of analysis. When this happens, serious methodological problems can be encountered (Herriott and Muse, 1973).

Coleman has made a number of contributions: a survey of methodological problems in sociological analysis including those encountered when trying to use social indicators for policy analysis (1969); an explication of the methodological foundations of policy research in the social sciences (1972a); and problems in using standardized tests to evaluate school performance (Coleman and Karweit, 1970). Rigsby and McDill have examined the conceptualization and measurement of adolescent peer influence processes (1972). Finally, Riley has reviewed a number of issues concerning the sources and types of sociological data (1964).

An NIE Strategy for Program Development

The measurement problem area is unique in that it cuts across all other problem areas, yet also stands apart as a discipline or sub-discipline in its own right with its own theory and methodology. Thus NIE faces the dilemma of choosing a centralized or de-centralized strategy in mounting initiatives to deal with the problems outlined in previous sections of this paper.

As has already been anticipated in earlier recommendations, a mixed strategy is advised, coinciding with the recommendations of the 1972 conference (Kooi, 1972). A completely decentralized approach is not desirable because "investigators working on substantive problems concentrate on those problems as such. They tend to employ current methods, even methods with known limitations, rather than turn aside to confront and resolve the methodological difficulties they meet" (Fiske, 1972). Furthermore, the use of common measures and common methodology among problem areas can be a powerful force toward reducing the fragmentation of education research and promoting the culumation of research knowledge. On the other hand, a completely centralized approach is not desirable either. Isolated measures have little meaning. They take on meaning as they are used to develop theories and models and to solve problems. This is the only way to estalbish the construct validity of measures.

Consequently, we recommend a mixed strategy in which certain functions and responsibilities are assigned to the various substantive programs within NIE while others are allocated to a central unit, and the two are tied together through form of matrix management.

Decentralized Functions

Each NIE program should include a compliment of measurement specialists. This group will often coincide with or overlap with those charged with evaluation functions within the program. They should be drawn not only from the tests and measurements field of educational psychology, but also from among measurement specialties in sociology, economics, and other disciplines.

Some measures tend to be unique to a problem (e.g. special instruments for bilingual populations) while others are common to many problems (e.g. turnover of personnel). While the Task Force on Measurement will attempt to identify and coordinate work on common measures, much of the work of instrument development, refinement, and validation must take place in the context of substantive research programs where their usefulness in theoretical models can be determined.

Theory and methodology of measurement can be handled best through a combination of intramural research and some targeting of field initiated research in the Office of Research Grants. If it is agreed that work in the measurement field should be an NIE priority and that we wish to stimulate an acceleration of work in the field, would be highly desirable

to identify a special Panel on Educational Measurement with its own funds. The Study Group working with this panel should work with the field to stimulate the flow of high quality proposals to the grant program.

Centralized Functions

One of the task forces within the Office of Research and Exploratory Studies should be made up of measurement specialists (possibly combined with concerns for methodology and secondary analysis, as seems to be the plan). This group should develop its own program of intramural and extramural research, concentrating on those problems that either cut across other programs or are not covered by other programs. These would include research on the effects of testing and other forms of measurement on individuals and systems, and work on new technologies such as item banks or computer testing. This staff should also serve as a resource for other programs in NIE when special needs arise. They would be the first group to whom the Director and Council would look when problems or inquiries regarding measurement arise. They would handle contacts and control correspondence with outside individuals making inquiries about measurement programs in NIE (with referral to more specific programs as appropriate).

We have also recommended that the NIE Library should have a measurement information specialist on its staff to assist NIE researchers in locating instruments, data banks, and the specialized literature of the measurement field.

Matrix Management

While it is possible to undertake most if not all of the work recommended in this paper through allocation of responsibilities in either the centralized or de-centralized mode, there remains a need to coordinate this work in order to maximise the synergistic effect. The fragmentation of effort has been one of the curses of educational research, and NIE needs to take special steps to avoid it. It is therefore proposed that a form of matrix management be utilized by forming an agency-wide Task Force on Measurement. This Task Force would be chaired by the director of the Task Force on Measurement, Methodology and Secondary Analysis in ORES and would include representation from the Study Group on Objectives, Measurement, and Evaluation of ORG, the Planning and Policy Analysis Unit of ORDR, the Educational Reference Division of OA, and measurement specialists in the line research units. This group should serve to coordinate work involving measurement in the several organizational units, promote the use of common measures where appropriate, cumulate and codify new knowledge as it emerges, develop standards for technical review of proposals, RFP's and products, and generally continue to build and refine an agency-wide strategy for the improvement of measurement for the various clients and purposes identified earlier. The effectiveness of such a group will be considerably enhanced if it has some funds at its disposal with which to support intramural research activities of its members.

Conclusion

This paper has covered a very diverse range of topics in a very broad field. Admittedly no one topic has been covered in the depth it deserves. However, a major purpose of the paper will have been served if the reader has gained a new conception of the range and complexity of the measurement field.

This work was originally undertaken because a number of reports prepared for the Planning Unit which preceded the establishment of NIE had recommended the development of instruments to measure a broader range of pupil outcomes. While the measurement of basic cognitive abilities is relatively well advanced, we do not have accurate and credible measures of other kinds of pupil performance that many consider important objectives of education, including problem-solving ability, moral values, social maturity, skill in interpersonal relationships, and other affective and higher order cognitive abilities.

While agreeing with the need for new pupil outcome measures, we have attempted to show that NIE should extend the range of its concern with measurement along a number of other dimensions as well.

(1) Our ability to measure characteristics of individuals is farther advanced than our ability to measure systems. Understanding the operations of systems is important both in its own right and in the contribution it can make to understanding individual growth and change.

(2) Similarly, psychometrics is a better developed field than the measurement sub-disciplines of sociology, political science, and other disciplines. As an inter-disciplinary problem area, educational R&D needs to include measurement research in all these fields, and NIE needs measurement specialists from each of them on its staff.

(3) Standardized tests represent only one way of collecting educational data. Support needs to be given to improvement of other data collection methods, including observation, questionnaires, interviews, administrative records, financial accounts, and other unobtrusive measures.

(4) The measurement needs of the research and development community are not coterminous with those of operating school systems. As an R&D agency NIE must contribute to the solution of measurement problems faced by researchers, developers, evaluators, and change agents as well as those of practitioners.

(5) While it is important to measure outcomes of education that correspond to explictly stated objectives, it is also important to detect and measure the unplanned and unintended consequences of educational programs.

(6) It is not enough to measure the outcomes of education at the individual or systems level. Research designs that treat schools as a "black box" are not likely to be useful. Our understanding of education

and the ability to devise solutions to problems depend on our ability to identify and measure inputs, contexts and processes related to those outcomes. Further, measures and the variables they represent cannot be neatly classified by analytic function; the same dimension might be an input, an output, or a context depending on the problem and the design.

(7) Above all, the importance of theory in deciding what ought to be measured needs to be recognized. It is not enough that technically correct instrument development techniques are used; there is a serious need to know more about what our instruments are measuring. A major effort should go into establishing the construct validity of measures. Wherever possible measures should be identified as part of larger systems of variables, theories, or models which seek to establish causal relationships.

BIBLIOGRAPHY

An asterisk (*) indicates that the work contains a compilation of measures, reviews of measures and/or information about measures used in a specific piece of research.

Anderson, James G. "Causal Models and Social Indicators: Toward the Development of Social System Models". American Sociological Review. 38:3 (June, 1973), pp. 285-301.

Anderson, Scarvia, and Samuel Messick. Social Competency in Young Children, ETS Report PR-73-9, under OCD Grant No. H-2993, March, 1973.

Anderson, Scarvia, Samuel Messick, and Nathaniel Hartshorne. Priorities and Directions for Research and Development Related to Measurement of Young Children, Report on Task 2 under OCD Grant No. H-2993 A/H/O, October, 1972.

* Beatty, Walcott H. (Chairman and Editor). Improving Educational Assessment and An Inventory of Affective Behavior, Prepared by the ASCD Commission on Assessment and Educational Outcomes. Washington, D.C.: National Education Association, 1969.

Beezer, Robert. "Research on the Construction of Psychometric Instruments", National Institute of education, no date.

Berg, Ivar. Education and Jobs: The Great Training Robbery. New York: Praeger Publishers, 1970.

Berliner, David C., and Leonard S. Cahen. "Trait-Treatment Interraction and Learning", in Fred N. Kerlinger (Ed.). Review of Research in Education 1, Itasca, Illinois: F.E. Peacock, 1973.

Blalock, Hubert M. "Comments on Coleman's Paper". in Robert Bierstedt (Ed.), A Design for Sociology: Scope, Objectives, and Methods. Philadelphia: The American Academy of Political and Social Science, 1969.

Block, James H. "Mastery Learning in the Classroom: An Overview of Recent Research". Santa Barbara: University of California at Santa Barbara, 1973.

* Bonjean, C.M., R.J. Hill, and S.D. McLemore, Sociological Measurement: An Inventory of Scales and Indices. San Francisco: Chandler Publishing Co., 1967.

Boocock, Sarane S. "The School as a Social Environment for Learning: Social Organization and Micro-Social Process in Education". Sociology of Education. 46:1, (Winter, 1973): pp. 15-50.

Boocock, Sarane, S. "Toward a Sociology of Learning: A Selective Review of Existing Research". Sociology of Education. 39:1, (Winter, 1966): pp. 1-45.

Brooks, Ralph M. "Social Planning and Societal Monitoring", in Leslie D. Wilcox et. al. Social Indicators and Societal Montioring: An Annotated Bibliography. Jossey-Bass/San Francisco, Elsevier: Washington, 1972: pp. 1-30.

* Buros, Oscar Krisen. The Seventh Mental Measurements Yearbook. Vols. I and II, 1972, Highland Park, New Jersey: The Gryphon Press.

Campbell, Angus and Philip E. Converse (Eds.). The Human Meaning of Social Change. New York: Russell Sage, 1972.

* CEMREL. Index of Tests and Measurements for Early Childhood Education. St. Louis: CEMREL, forthcoming.

Clark, Burton R., (Chairman). "Sociology and the Study of Education". Report of a Planning Conference for the NIE Planning Unit, July 30-31, 1971. Report C106. Washington, D.C.: NIE, 1972.

Cobern, Morris, Claude Salem, and Selma Mushkin. Indicators of Educational Outcome. Washington, D.C.: Government Printing Office, 1973.

Cohen, Elizabeth G. "Sociology and the Classroom: Setting the Conditions for Teacher-Student Interaction", Review of Educational Research. 42:4. (Fall, 1972): pp. 441-452.

Coleman, James S. "Policy Research in the Social Sciences", General Learning Corporation, 1972. (a)

Coleman, James S. "How do the Young Become Adults?", Review of Educational Research, 42:4. (Fall, 1972): (b) pp. 431-439.

Coleman, James S. Resources for Social Change: Race in the United States. New York: Wiley-Interscience, 1971.

Coleman, James S., "The Methods of Sociology", In Robert Bierstedt (Ed.), A Design for Sociology: Scope, Objectives and Methods. Philadelphia: The American Academy of Political and Social Science, 1969: pp. 86-114.

Coleman, James S., and Nancy L. Karweit. Measures of School Performance. R-488-RC. Santa Monica: The Rand Corp., July, 1970.

Coleman, James S., et. al., Equality of Educational Opportunity. Washington, D.C.: Office of Education, Department of Health, Education and Welfare, Government Printing Office, 1966.

* Corwin, Ronald G. Militant Professionalism: A Study of Staff Conflicts in High Schools. New York: Appleton-Century, Crofts, 1970.

Crandall, Rick. "The Measurement of Self-Esteem and Related Constructs". Measures of Social Psychological Attitudes. (1973), Ann Arbor: University of Michigan.

Cronbach, Lee J. "Test Validation", in Robert L. Thorndike (Ed.). Educational Measurement. 2nd ed. Washington, D.C.: American Council on Education, 1971.

Cronbach, Lee J. Essentials of Psychological Testing, third edition. New York: Harper and Row, 1970.

Dreeben, Robert. "The Contribution of Schooling to the Learning of Norms". Harvard Education Review, 37:2, 211-237 (Spring, 1967): pp. 23-49. (Reprinted in Socialization and Schools, Harvard Education Review, 1968).

Dreeben, Robert. On What is Learned in School. Reading, Massachusetts: Addison-Wesley, 1968.

Duncan, Otis Dudley. "Social Stratification and Mobility: Problems in the Measurement of Trend", in Sheldon, Eleanor Bernert and Wilbert E. Moore (Eds.). Indicators of Social Change: Concepts and Measurements. (1968), New York: Russell Sage Foundation pp. 675-719.

Ebel, Robert L. "The Future of Measurements of Abilities II". Educational Researcher. (March, 1973): pp. 5-12.

Educational Testing Service Assessment in a Pluralistic Society. Proceedings of the 1972 Invitational Conference on Testing Problems, Princeton, New Jersey, 1973.

Etzioni, Amitai. An NIE Strategy Paper. Washington, D.C.: NIE, 1972,

Fennessey, James. "Using Achievement Growth to Analyze Educational Programs". Report No. 151. Center for Social Organization of Schools, Johns Hopkins University, 1973.

Ferriss, Abbott L. "Monitoring and Interpreting Turning Points in Educational Indicators", Proceedings of the Social Statistics Section(1972), Washington, D.C.: American Statistical Association; pp. 60-65.

Ferriss, Abbott L. Indicators of Trends in American Education.New York: Russell Sage Foundation, 1969.

Fiske, Donald W. "Draft Statement for a Program Area on Measurement: Theory and Methodology for Measurement and Evaluation in Educational Research". unpublished.

Glaser, "Instructional Technology and the Measurement of Learning Outcomes". American Psychologist.(1963): pp. 510-522.

Goslin, David A., (Ed.). Handbook of Socialization Theory and Research. Chicago: Rand McNally, 1969.

Gross, Neal, Joseph B. Giacquinta, and Marilyn Bernstein. Implementing Organizational Innovations: A Sociological Analysis of Planned Educational Change. New York: Basic Books, 1971.

* Grotberg, Edith. "Early Childhood Research and Development Needs and Gaps in Federally Funded Intervention Studies within a Longitudinal Framework". Washington, D.C.: Social Research Group, GWU, 1972.

Grotberg, Edith, and Ellen Searcy. "A Statement and Working Paper on Longitudinal/Intervention Research". Washington, D.C.: Social Research Group, George Washington University, 1972.

* Guthrie, J.W. "A Survey of School Effectiveness Studies". in Do Teachers Make A Difference? Washington, D.C.: Government Printing Office, 1970.

Health, Education and Welfare, U.S. Department Of. Toward A Social Report. Washington, D.C.: Government Printing Office, 1969.

Herriott, Robert E., and Donald N. Muse, "Methodological Issues in the Study of School Effects", in Fred N. Kerlinger (Ed.). Review of Research in Education 1. Itasca, Illinois: F. E. Peacock, 1973: pp. 209-236.

Herriott, Robert E., and Benjamin J. Hodgkins. The Environment of Schooling: Formal Education as an Open System. Englewood Cliffs, New Jersey: Prentice Hall, 1973.

Hively, Wells, et. al. Domain-Referenced Curriculum Evaluation: A Technical Handbook and a Case Study from the Minnemast Project. Los Angeles: CSE Monograph Series in Evaluation, Vol. 1, Center for the Study of Evaluation, UCLA, 1973.

Hoepfner, Ralph. "Published Tests and the Needs of Educational Accountabilty", Educational and Psychological Measurement, forthcoming (1974).

Hoepfner, Ralph, Paul A. Bradley, Stephen P. Klein, and Marvin C. Alkin. CSE/Elementary School Evaluation Kit: Needs Assessment. Boston: Allyn and Bacon, Inc., 1972.

* Hoepfner, Ralph, et. al. CSE Elementary School Test Evaluations. Center for the Study of Evaluation. Los Angeles: UCLA Graduate School of Education, 1970.

* Hoepfner, Ralph, et. al. CSE-ECRC Preschool/Kindergarten Test Evaluations. Los Angeles: UCLA Graducate School of Education, 1971.

* Hoepfner, Ralph, et. al. CSE-RBS Test Evaluations: Tests of Higher-Order Cognitive, Affective, and Interpersonal Skills. Los Angeles: Center for the Study of Evaluation, Graduate School of Education, UCLA, 1972.

Hyman, Herbert H. Secondary Analysis of Sample Surveys: Principles, Procedures, and Potentialities. New York: John Wiley & Sons, Inc, 1972.

Inkeles, Alex. "Social Structure and the Socialization of Competence". Harvard Educational Review. 36:3, 1966.

Inkeles, Alex. "Social Structure and Socialization", in Goslin, David A. (Ed.). Handbook of Socialization Theory and Research. Chicago: Rand McNally, 1969: pp. 615-632.

Jencks, Christopher. Inequality: A Reassessment of the Effects of Family and Schooling in America. New York: Basic Books, 1972.

Joyce, Bruce R. Alternative Models of Elementary Education. Waltham, Massachusetts: Blaisdell Publishing Co., a Division of Ginn & Co. 1969.

Kirkland, Marjorie C. "The Effects of Tests on Students and Schools". Review of Educational Research. 41:4 (October 1971): pp. 303-350.

Kluckhohn, Florence R. "Dominant and Variant Value Orientations", in Cylde Kluckhohn and Henry A. Murray (Eds.). Personality in Nature, Society and Culture. New York: Alfred A. Knopf (1953): pp. 342-357.

Kooi, Beverly. "New Measures for Education: A Proposed Agenda for NIE". 1972, unpublished.

Kooi, Beverly. Program Planning for the National Institute of Education: A Summary of Four R&D Analyses. Washington, D.C.: NIE, June 1972.

Kooi, Beverly, et. al. A Research and Development Agenda for the National Institute of Education. Washington, D.C.: NIE, July 1972.

Krantz, R. Luce, Patrick Suppes, and A. Tversky. Foundations of Measurement. Vols. 1 and 2. Academic Press, 1972.

Krathwohl, David R., and David A. Payne. "Defining and Assessing Educational Objectives", in Robert L. Thorndike (Ed.). Educational Measurement. 2nd ed. Washington, D.C.: American Council on Education, 1971: p.

* Lake, Dale G., Matthew B. Miles, and Ralph Earle, Jr. Measuring Human Behavior. New York: Teachers College Press, 1973.

Land, Kenneth C. "Social Indicator Models: An Overview". Paper delivered at AAAS. December 1972.

* Langenfeld, James. "Empirical Findings on Self-Esteem: A Selected Survey". Washington, D.C.: Public Services Laboratory, Georgetown University, 1972.

Larson, Robert, Wayne Martin, Donald Searls, Susan Sherman, Todd Rogers and David Wright. "A Look at the Analysis of National Assessment Data", Paper presented by J. Stanley Ahmann at the Invitational Conference on the Occasion of the Dedication of the E. F. Lindquist Center for Measurement, University of Iowa, Iowa City, Iowa, April 6-7, 1973, National Assessment of Educational Progress, Denver, 1973.

* Lawrence, Benjamin, G. Weathersby, and V. W. Patterson (Eds.). Outputs of Higher Education: Their Identification, Measurement, and Evaluation. Boulder, Colorado: Western Interstate Commissioner for Higher Education, 1970.

* Lazar, Joyce B. "A Preliminary Report on the Present Status and Future Needs in Longitudinal Studies in Early Childhood Research and Development". Washington, D.C.: The Social Research Group, George Washington University, 1972.

Levien, Roger E. National Institute of Education: Preliminary Plan for the Proposed Institute. R-657-HEW. Santa Monica, The Rand Corp., February, 1971.

Levien, Henry M. "A Conceptual Framework for Accountability in Education". Occasional paper: Stanford University, September, 1972.

* MacDonald, A. P., Jr. "Internal-External Locus of Control". Measures of Social Psychological Attitudes, Ann Arbor: University of Michigan, 1973.

Madaus, George, and Richard F. Elmore. "Allocation of Federal Compensatory Education Funds on the Basis of Pupil Achievement Test Performance". Statement submitted to the General Education Subcommittee of the Committee on Education and Labor: Washington, D.C.: U.S. House of Representatives, June 26, 1973.

Mason, Ward S. A Method of Scaling Cultural Orientations". PhD disseration. Cambridge, Massachusetts: Department of Social Relations, Harvard University, 1952.

Mason, Ward S. "Progress Report on Proposed Programs for NIE Focusing on New Measures In Education". Washington, D.C.: NIE, 1972.

McClelland, David C. "Testing for Competence Rather and for 'Intelligence'". American Psychologist. January, 1973: pp. 1-14.

Messick, Samuel, and Scarvia Anderson. "Educational Testing, Individual Development, and Social Responsibility". The Counseling Psychologist. 2:2, (1970): pp. 80-88.

Mushkin, Selma J. "Performance Toward What Result: An Examination of Some Problems in Outcome Measurement". Prepared for the National Conference on Performance Contracting, Belmont House, Elkridge, Maryland. Washington, D.C.: Public Services Laboratory, Georgetown University, 1971.

Mushkin, Selma J. "The 'SIR' Adjusted Index of Educational Achievement". Public Services Laboratory, Georgetown University, Washington, D.C." Office of Education Contract OEC-0-70-4454(521), no date.

Mushkin, Selma. Educational Outcomes: An Exploratory Review of Concepts and their Policy Application, A report to NCES. Washington, D.C.: Public Services Laboratory, Georgetown University, April, 1972.

Mushkin, Selma. "National Assessment and Social Indicators". DHEW Pub. No. (OE)73-11111. National Center for Educational Statistics. Washington, D.C.: Government Printing Office, 1973.

National Commission on Technology, Automation, and Economic Progress. Technology and the American Economy. Washington, D.C.: 1966.

National Institute of Education. Partial Bibliography of Reports Related to the National Institute of Education. Washington, D.C.: Office of Public Information, 1973.

Nixon, Richard M. "Message from the President of the United States on Educational Reform". Document No. 91-267, Washington, D.C.: 91st Congress, House of Representatives, March 3, 1970.

Payne, David A., and Richard W. Watkins. "The Inter-Association Council on Test Reviewing: Alpha and Omega". Educational Researcher, July, 1973, 18-20.

Popham, W. James. "California's Precedent-Setting Teacher Evaluation Law". Educational Researcher. July 1972: pp. 13-15.

Popham, W. James, and T.R. Husek. "Implications of Criterion Referenced Measurement", in Criterion Referenced Measurement, W.J. Popham (Ed.) Cliffs, N.J.: Educational Technology Publishers, 1971.

* Price, James. Handbook of Organizational Measurement. Homewood, Ill.: Richard D. Irwin, Inc., 1973.

* Rigsby, Leo C., and Edward L. McDill. "Adolescent Peer Influence Processes: Conceptualization and Measurement". Social Science Research. 1:3 (September 1972): pp. 305-321.

Riley, Matilda White. "Sources and Types of Sociological Data". Handbook of Modern Sociology, Robert E.L. Faris (Ed.). Chicago: Rand McNally & Co., 1964, 978-1026.

* Robinson, John P., Robert Anthanasiou, and Kendra B. Head. Measures of Occupational Attitudes and Occupational Characteristics. University of Michigan: Survey Research Center, Institute for Social Research, 1969.

* Robinson, John P., and Jerrold G. Rusk, and Kendra B. Head. Measures of Political Attitudes. University of Michigan: Institute for Social Research, 1968.

Rossi, Peter H., and Walter Williams (Eds.). Evaluating Social Programs. New York: Seminar Press, 1972.

Scriven, Michael. "Prose and Cons about Goal-Free Evaluation". Evaluation Comment, 3:2 (December, 1972): pp. 1-4.

* Shaw, M.E., and J.M. Wright. Scales for the Measurement of Attitudes. New New York: McGraw-Hill, 1967.

Sheldon, Eleanor Bernert, and Howard E. Freeman. "Notes on Social Indicators: Promises and Potential". Policy Sciences 1 (1970): pp. 96-111.

Sheldon, Elearnor Bernert, and Kenneth C. Land. "Social Reporting for the 1970's". Policy Sciences 3 (1972): pp. 137-151.

Sheldon, Eleanor Bernert, and Wilbert E. Moore, (Eds.) Indicators of Social New York: Russell Sage Foundation, 1968.

* Simon, Anita and E. Boyer (Eds.). Mirrors for Behavior: An Anthology of Observation Instruments Vols. I-VI, Philadelphia: Research for Better Schools, Inc., 1967.

Simon, Anita and E. Boyer (Eds.). Mirrors for Behavior: An Anthology of Observation Instruments Continued Vols. VII-XV, Philadelphia: Research for Better Schools, Inc., 1970.

* Solomon, Warren, Daniel Ferritor, Joseph Haenn, Edwin Myers. "The Development, Use, and Importance of Instruments that Validly and Reliably Assess the Degree to which Experimental Programs are Implemented". St. Ann, Mo.: CEMREL, Inc., 1973.

Spady, William G. "The Impact of School Resources on Students" . in Fred N. Kerlinger (Ed.), Review of Research in Education 1, Itasca, Ill.: F. E. Peacock, 1973: pp. 135-177.

Stake, Robert E. "School Accountability Laws." Evaluation Comment, Vo. 4, No. 1, February, 1973: p. 1.

Suchman, Edward A. "Evaluating Educational Program". in Francis G. Caro (Ed.), Readings in Evaluation Research, New York: Russell Sage Foundation, 1971: pp. 43-48.

Suchman, Edward A. Evaluative Research: Principles and Practice in Public Service and Social Action Programs, New York: Russell Sage Foundation, 1967.

* Taylor, James C. and David G. Bowers. Survey of Organizations: A Machine Scored Standardized Questionnaire Instrument, Ann Arbor: Center for Research on Utilization of Scientific Knowledge Institute for Social Research, University of Michigan, 1972.

Thorndike, Robert L. (Ed.). Educational Measurement. 2nd ed. Washington, D.C.: American Council on Education, 1971.

Trent, James W. et. al. An Analytical Review of Longitudinal and Related Studies as they Apply to the Educational Process, 5 vols. Los Angeles: Center for the Study of Evaluation, University of California, 1972-73.

* Walizer, M.W., Robert E. Herriott. The Impact of College on Students' Competence to Function in a Learning Society, Report No. 47, Iowa City: Research and Development Division, American College Testing Program, 1971.

* Wilcox, Leslie D., Ralph M. Brooks, George M. Beal, and Gerald E. Klonglan. Social Indicators and Societal Monitoring: An Annotated Bibliography, San Francisco - Washington: Jossey-Bass/Elsevier, 1972.

Williams, Charles. Anticipating Educational Issues Over the Next Two Decades: An Overview Report on Trends Analysis, Menlo Park, Calif: Educational Policy Research Center, Stanford Research Institute, 1973.

Williams, R.L. "Black Pride, Academic Relevance and Individual Achievement". The Counseling Psychologist, 2:1. 1972: pp. 18-22.

Wholey, Joseph S., John W. Scanlon, Hugh G. Duggy, James S. Fukumoto, and Leona M. Vogt. Federal Evaluation Policy: Analyzing the Effects of Public Programs, Washington, D.C.: The Urban Institute, 1970.

Womer, Frank B. Measurement in Education: National Assessment Says. East Lansing: National Council on Measurement in Education, October, 1970.