DOCUMENT RESUME

ED 275 736 TM 860 650

AUTHOR	Baker, Eva L.
TITLE	NIE Planning Grant for Center on Student Testing, Evaluation, and Standards. Performance Report.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	Aug 85
NOTE	47p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Educational Assessment; *Educational Planning; Educational Quality; Educational Research; *Educational Testing; Grants; Models; *Program Design; *Program Proposals; Research and Development; Research Needs; *Standards; *Student Evaluation
IDENTIFIERS	Center for the Study of Evaluation CA; National Institute of Education

ABSTRACT

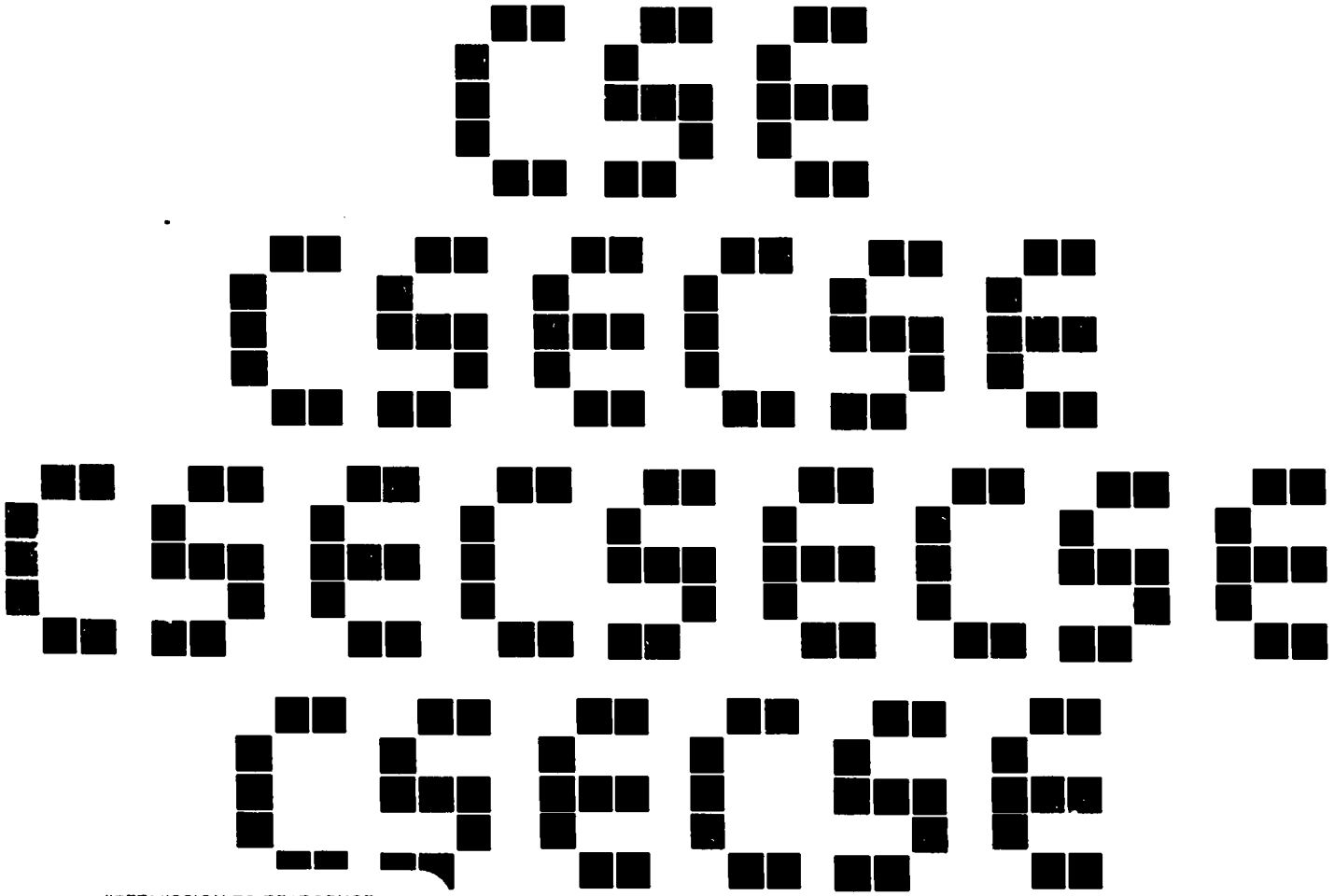
This document summarizes the activities conducted under the National Institute of Education (NIE) planning grant for the new Center on Student Testing, Evaluation, and Standards and the conceptualizations that emerged from these activities. Chapter 1 presents a summary of the planning activities actually conducted under the award, including particular problems and successes and a list of participants and their affiliations. Chapter 2 provides a technical report on the Research and Development mission for a Center on Student Testing, Evaluation and Standards, including a brief review of the literature, analysis of problems in practice, guiding themes for the research agenda and effective strategies for conducting the research. Chapter 3 is a futures paper which summarizes in nontechnical language the proposed Center's mission, long rangs plans, and objectives. A four-page list of references concludes the document. (Author/JAZ)

Center for the Study of Evaluation

UCLA Graduate School of Education
Los Angeles, California 90024

PERFORMANCE REPORT

NIE Planning Grant for Center on Student Testing, Evaluation, and Standards



PERFORMANCE REPORT

NIE Planning Grant for Center on Student Testing, Evaluation, and Standards

Eva L. Baker
Principal Investigator

Center for the Study of Evaluation
University of California, Los Angeles
August 1985

TABLE OF CONTENTS

PERFORMANCE REPORT

NIE Planning Grant for Center on
Student Testing, Evaluation, and Standards

Introduction

This document summarizes the activities conducted under the NIE planning grant for the new Center on Student Testing, Evaluation, and Standards and the conceptualizations that emerged from these activities. As required by the grant, the document is organized as follows:

- Chapter One presents a summary of the planning activities actually conducted under the award, including particular problems and successes and a list of participants and their affiliations.

- Chapter Two provides a technical report on the R&D mission for a Center on Student Testing, Evaluation and Standards, including a brief review of the literature, analysis of problems in practice, guiding themes for the research agenda and effective strategies for conducting the research.

- Chapter Three is a futures paper which summarizes in nontechnical language the proposed Center's mission, long range plans, and objectives.

CHAPTER ONE: Summary of Planning Activities

Actual activities conducted under the planning grant varied from those initially planned due to the delay in the competition and the additional guidelines provided by the NIE for the mission area. As a result of these NIE instituted changes, the proposed mission needed to be amended from the planning grant proposal. Instead of devoting the planning process to soliciting review of the mission proposed in the planning grant, then, substantial time and effort had to be allocated to the mission amendment. While the amendment abbreviated the wide and repeated review process initially intended, the planning process nonetheless gathered wide input from a number of stakeholders with interests in educational testing and evaluation. In the sections which follow, we provide a chronology of the planning process, describing particular problems and successes as they occurred, and a losting of the participants in the process. Please note that these planning activities benefitted from University contributed resources as well as from funds granted by the NIE.

Chronology of Planning Activities

Planning was comprised of five major activities. These included collaboration with members of the National Faculty to get their feedback on priority areas for the research mission and effective strategies for the conduct of R&D; mission and research planning by members of the Research Council and participating faculty and staff; review of mission and research plans by noted researchers and practitioners followed by revisions as necessary; and planning for collaboration with other laboratories, centers, and state and local practitioners and policymakers. Specific activities within each of these areas are described below.

Collaboration with members of the National Faculty to get feedback on priority areas for research. In the planning proposal, we advanced the idea of a National Faculty of interested practitioners and policymakers who would collaborate with us during all stages of the research process, from planning through dissemination; and in fact conversations with some of these individuals influenced the perspectives described in the planning proposal. Once the planning award was granted, an initial task was to get systematic and specific feedback on the mission and research we proposed from a National Faculty composed of members representing the full range of interests, including teachers, administrators, school board members, state and local superintendents, state and local directors of research and evaluation, military trainers, and test publishers. A meeting of representatives from each of these groups was convened at the annual meeting of the American Educational Research Association in April, 1985 as a first step in this feedback process. The proposed mission was discussed as were ideas and specific projects. National Faculty members were given copies of the proposed mission and questionnaires for soliciting their feedback on the following issues:

- Overall importance of the problems addressed by the mission;
- Additional problems that ought to be addressed;
- Probable effectiveness of strategies for conducting RDD&E;

- Additional strategies that should be included;

- The single most important objective that should be addressed by a national center on student testing, evaluation, and standards;

- The importance of the issues emphasized in each of the proposed research programs, and the need for additions, deletions and/or modifications;

- Reactions to each of the potential research initiatives in terms of its relevance to the mission; importance of the problem addressed; potential impact on policy/practice; likelihood of success.

In addition to providing their own reactions, members also were asked to solicit the reactions of their peers and to report to us accordingly. They were asked to provide such feedback by May 31.

This feedback was summarized quantitatively and qualitatively for the June deliberations of the Research Council (see below). Reactions from the National Faculty were uniformly positive with regard to the importance of the mission and the probable effectiveness of proposed RDD&E strategies. They highly rated the proposed research programs and were generally favorable toward all of the research initiatives. Respondents appeared most positive about R&D related to the use of testing for instructional improvement and about gaining knowledge about how to deal with practical problems.

Members of the National Faculty generally were enthusiastic about the opportunity to offer their views and to collaborate in research planning. In some cases, particularly at the highest administrative levels, enthusiasm exceeded available time to carefully review proposal documents and to respond in depth to them within time constraints. In these cases, feedback was more informal, through personal interaction and conversation. Other sources of feedback included informal meetings at conferences attended during the planning period, e.g., the ECS conference at Boulder provided an opportunity to meet with many state level decisionmakers, a College Board sponsored equity conference enabled meetings with local and state administrators and subject matter experts in a range of disciplines.

Mission and research project planning by the Research Council. As proposed in the planning proposal, a Research Council composed of Center directors, program directors, and representatives of each collaborating institution (University of Illinois, University of Colorado, National Opinion Research Center at the University of Chicago, and Educational Testing Service) was to be central in clarifying the Center's mision and in making decisions about what projects to fund during the first proposal period. A meeting of this group was convened at UCLA on March 14-15, 1985 to reach consensus on directions for the mission, program organization, and to hear presentations on high priority research initiatives that might be funded; management structure for the new NIE Center and schedules for completing and reviewing drafts of the proposal were also discussed.

The Research Council was scheduled to meet with the National Faculty during the annual meeting of the American Educational Research Association and subsequently to make final decisions about the mission and research projects. While members of the Research Council did meet with the National Faculty about general mission directions, decisions about specific projects to be funded and their planning were postponed until after Secretary Bennett reviewed the mission area and published suggested modifications.

Subsequent to Secretary Bennett's announced modifications, Research Council members as well as interested faculty and staff were requested to submit additional ideas for research projects, including a description of the problem to be addressed, its significance in relation to mission, proposed methodology, and budget requirements. These proposals were presented to the Research Council at a meeting held on June 4-5, 1985 at UCLA.

At this meeting, the Research Council considered the reactions of the National Faculty to the proposed mission and the modifications suggested by Secretary Bennett. Based on their deliberations, they reached consensus on the revised mission which was to guide the research programs, including major themes and Center objectives. After agreeing on the mission, the Research Council considered each proposed research project in relation to the Center mission and objectives, its potential contribution to the improvement of practice and to the development of theory and understanding of fundamental issues, its intellectual rigor, and its possible interrelationships with other proposed projects. Discussion focused on project options and modifications which would increase the coherence of the proposed projects within and across programs and/or which might be most cost effective in producing a balanced overall program of research. Based on these discussions, the directors of the proposed Center made recommendations for the initial slate of projects to be funded and the resources to be allocated to each. The Research Council concurred with the Directors' recommendations. After reaching agreement on projects, team meetings by program were held to refine key themes and objectives for each program, to specify program study teams for future projects, and to discuss inter-relationships between projects and ways to facilitate aggregation of findings. Responsibilities and schedules for producing the proposal were then reiterated. Drafts sections of the proposal were to be completed by June 28 and subsequently reviewed both by members of the Research Council and by external reviewers and then revised as necessary.

Review of proposed mission and research programs. As drafts of the mission and strategy, operational plans for research and institutional functions and institutional capacity sections were completed, they were reviewed first by the Center directors and members of the Research Council and modified as necessary to increase the coherence of the proposed work and its methodological rigor. After initial review and revision process, drafts of the entire proposal package were reviewed thoroughly by both educational practitioners (Dr. Steven Winkel from Montgomery County Schools and Dr. Lynn Winters from Palos Verdes Unified Schools) and by noted researchers in the field of testing and evaluation (Dr. C. Robert Pace and Dr. Samuel Messick). These individuals were asked in particular to comment on the coherence of the mission, the integration, significance

and methodological strength of proposed research programs and to offer suggestions for improvement. Subsequent to these reviews, the proposal document was revised and the final document produced.

Planning for collaboration. Concurrent with the activities described above, contacts were made with competitors for research centers which were likely to have overlapping interests with a center on student testing, evaluation and standards. These included the centers on writing, learning, effective elementary schools, effective secondary schools, state and local policy, post-secondary teaching and learning. Principal investigators and other key personnel were contacted at Chicago, Johns Hopkins, Stanford, Pittsburgh, Teachers College, Michigan State, Wisconsin, Florida State, Rutgers, Hartford, and Berkeley to discuss potential areas of mutual concern and to agree, if successful in the competition, to future meetings devoted to planning collaborative ventures. Two ideas for collaboration which evoked considerable interest were participation in study groups aimed at important problems in educational policy and/or practice (e.g., quality indicators for the precollegiate and post-secondary levels); and sponsoring joint conferences exploring methodological issues in conducting research and evaluation in a specific topic area (e.g., effective schools).

Participants in the Planning Process

The activities describe above involved individuals from the researcher, practitioner, and policymaking communities. These individuals included:

Marvin Alkin, University of California, Los Angeles
Gordon Ambach, Commissioner of Education, New York
Ernest Anastasio, EDUCOM
Josie Bain, Los Angeles Unified School District
Eva Baker, University of California, Los Angeles
Adrianne Bank, University of California, Los Angeles
Darrell Bock, University of Chicago
James Burry, University of California, Los Angeles
Leigh Burstein, University of California, Los Agneles
Beverly Cabello, University of California, Los Angeles
Dale Carlson, California State Dept. of Education
James Catterall, University of California, Los Angeles
William Cody, Supt. of Schools, Montgomery County
David Cohen, University of California, Los Angeles
Elaine Craig, University of California, Los Angeles
Phil Curtis, University of California, Los Angels,
Don Dorr-Bremme, University of California, Los Angeles
Walter Feurzeig, Bolt, Beranek & Newman, Inc.
Steve Frankel, Montgomery County Public Schools
Calvin Frazier, Commissioner of Education, Colorado
Howard Freeman, University of California, Los Angeles
Gene Glass, University of Colorado
Wayne Gordon, University of California, Los Angeles
William Harris, Educational Testing Service
Joan Herman, University of California, Los Angeles
Ernest House, University of Illinois
Pete Idstein, Christina Unified School District, Delaware

Jim Johnson, University of California, Los Angeles
Mary Johnson, Dept. of Defense Dependent Schools
Tom Kerins, Illinois State Board of Education
Ward Keesling, Advanced Technology, Inc.
Harold Levine, University of California, Los Angeles
Robert Linn, University of Illinois
David McArthur, University of California, Los Angeles
Bernard McKenna, (NEA) National Education Association
Joyce McLarty, Tennessee State Department of Education
James Mecklenburger, National School Boards Association
Samuel Messick, Educational Testing Service
Jason Millman, Cornell University
Bengt Muthen, University of California, Los Angeles
James Olsen, WICAT
Robert Pace, University of California, Los Angeles
Sharon Robinson, National Education Association
Edward Roeber, State Department of Education
Gila Saks, University of California, Los Angeles
Francisco D. Sanchez Jr., Supt. of Schools - Albuquerque (retired)
Tom Satterfield, Deputy State Supt., Mississippi Dept. Ed.
Geoffrey Saxe,University of California, Los Angeles
Richard Shavelson, University of California, Los Angeles
Lorreta Shepard, University of Colorado
Kenneth Sirotnik, University of California, Los Angeles
Marshall Smith, University of Wisconsin
Mary Lee Smith, University of Colorado
Harris Sokoloff, University of Pennsylvania
Eliott Soloway, Yale University
Robert Stake, University of Illinois
Floraline Stevens, Los Angeles Unified School District
Ron Tarr, U.S. Army Infantry School
James Ward, American Federation of Teachers
William Ward, Educational Testing Service
Noreen Webb, University of California, Los Angeles
Richard Williams, University of California, Los Angeles
Lynn Winters, Palos Verdes School District
Merlin Wittrock, University of California, Los Angeles

CHAPTER TWO: Technical Report on R&D Mission

This report outlines a mission for an R&D Center on Student Testing, Evaluation, and Standards. It begins with a brief review of the literature, highlighting the authors' perceptions of current important research directions for testing, evaluation, and standards. Next, a synopsis of current problems is presented, followed by a conceptual framework for conducting R&D on these problems. The report concludes with a summary of the R&D objectives inherent in the framework.

The View From 1985

During the last 15 years, testing and evaluation scholars and practitioners have learned a prodigious amount. They have redefined evaluation impact so that it is now much more than a simple technical issue. They have proposed models, approaches, analyses, and solutions to recurrent problems. During this period, too, evaluation and testing have come to play much larger roles in public policy.

In testing, technical developments in item response theory (IRT) (e.g., Bock & Aitkin, 1981; Lord, 1980) have provided a new and powerful means of attacking previously intractable problems such as detecting biased test items (e.g., Shepard et al, 1981), constructing and equitably scoring computerized adaptive tests (Green, Bock, Humphreys, Linn, & Reckase, 1984), and creating and calibrating of multipurpose item banks for the effective assessment of individual students as well as instructional programs (Bock, Mislevy, & Woodson, 1982). The conception of testing has evolved from an unquestioned dependence on differentiation among students, to an emphasis on content encouraged by the criterion-referenced testing movement that followed Glaser's (1963) landmark paper. Concurrent with the renewed emphasis on content has been the forging of a promising linkage between psychometrics and cognitive psychology (e.g., Brown & Burton, 1978; Curtis & Glaser, 1983; Tatsuoka & Tatsuoka, 1982). Together, these achievements represent the first step toward an integration of testing and instruction (See, for example the Special Issue of Journal of Educational Measurement, 1983).

In evaluation, simple linear models of evaluation, thought to mirror a linear pattern of needs identification, planning, implementation, and evaluation (see e.g., Alkin, 1969; Stufflebeam et al, 1971), have been replaced by analyses that recognize the complex interactions of technical, social, structural, and political environments (e.g., Bank & Williams, 1984a, 1984b; Cronbach, 1982; Cronbach et al 1980; Patton, 1978; House, 1977; Weiss, 1972). From simple, controlled studies of outcomes, design and data collection have been augmented to include studies of how policy goals, implementation and multifaceted information interact (e.g., Berman & McLaughlin, 1977; Cook & Campbell, 1979; Stake, 1978). Studies of evaluation have been enlarged to reflect a concern that the results be used by a range of decisionmakers (e.g., Alkin, Daillak, & White, 1979; Bryk, 1983; Reisner et al, 1982).

The mission of evaluation now goes beyond the analysis and judgment of particular programs (Cronbach et al, 1980). Its scope has expanded to include consideration of how integrated evaluation systems work and how they can be improved. To that end, evaluation information--and the testing programs that support it--should help to clarify standards; and the whole process should serve to target resources and to stimulate effort in areas of critical need.

Our vantage point suggests that the models driving evaluation must be formative (Scriven, 1967) and that they must attend to the shift in emphasis to state and local initiative and responsibility. Educational interventions are rarely treatments in the traditional research sense (Burstein & Guiton, 1984); rather they are subject to a range of local adaptations, surprise turns, and altered expectations. As a result, formative evaluation requires a thorough understanding of the context in which evaluation findings are developed and are expected to be implemented; of the social, structural and political contexts in which education resides, and of the pragmatics of life in the schools (Baker, 1981; Sirotnik, Burstein and Thomas, 1983).

The efforts of the proposed CSTES must be guided by the following questions: What test and other information creates the potential for improvement? How should the quality of information be judged and improved--that is, how can the information be made more credible, valid, and ultimately useful? (Cronbach, 1982; see also our expanded view in Appendix 4.) The characteristics of useful information depend upon one's perspective (Dorr-Bremme, 1983; Sirotnik et al, 1983). To be useful to students and teachers, information should probably be very specific, should be carefully timed, and should be presented in a way that takes into account the limits of what can be productively absorbed. The way in which information is conveyed and displayed is also important (Sirotnik & Burstein, 1984). For instance, school and district managers may require detailed analyses of educational services and policies rather than detailed outcome information (Burstein, 1981); higher-level policymakers may demand comparability of information; and the public at large probably prefers credible generalizations without a lot of detail and backup evidence (Smith, 1984).

The proposed CSTES must also be sensitive to possible conflicts between information that will contribute to the top-down demand for broad-level accountability (to improve management and to elevate standards of excellence) and the bottom-up demand for adaptive, sensitive information that will be useful at the local level (Baker, 1983). These two sets of demands push in different and not-totally- compatible directions. Some of the tensions are obvious. A testing and evaluation system whose purpose is instructional improvement requires information which is based on local expectations and resources, which is adaptive to unplanned changes, and which is timed so that options can be assessed and selected. But external requirements pull in the direction of comparable, more uniform designs for information.

Both top-down and bottom-up points of view would be better served by an expanded view of standards of educational quality. These data requirements extend beyond student attainment of particular subject matters or basic skills to information about student learning processes, educational services, instructional processes, and important contextual factors. Such a data base approach implies that information needs will be driven selectively by the pragmatics of the environment.

What is the potential role of disciplined inquiry (Howe, 1984) in addressing the competing expectations for information, and what insights can it provide into the concerns for quality expressed in A Nation at Risk (1983) and in other prestigious reports (Goodlad, 1983; Boyer, 1983; Sizer, 1984)? How can the new standards articulated by legislatures and by local school boards be connected to a broadened view of educational quality? What can science, research and development, and conceptual analysis contribute to productive educational reform, and what are their limitations? One important function of the CSTES is to answer questions such as these.

This view from 1985 reflects our perceptions of the current important research directions for testing, evaluation, and standards. Guiding these perceptions are several global beliefs:

- Testing, evaluation, and standard setting can contribute to improving the quality of education. Tests -- when they are well conceived, constructed, administered, and analyzed -- can provide valuable insights into how individuals and classes of students are learning; they can help guide teaching, administration, and policy-making within our educational institutions. Evaluations of programs -- especially when they are seen as improvement oriented, locally useful, and iterative -- can help to guide the reallocation of resources, the modification and improvement of activities, and the retraining of personnel. Standards -- set with due attention both to what is desirable and to what is feasible at the state and local levels -- can help to focus attention and promote accountability for educational improvement.

- Testing and evaluation are important tools for promoting educational equity. Tests, when they are sensitive to individual differences and preferences in learning styles, provide a powerful means for diagnosing students' unique needs and providing effective instruction for all students. Furthermore, tests, when they match classroom instruction, can provide fair and equitable measures of student progress, measures which focus on learning accomplishments rather than background characteristics. Achievement measures as well as measures of educational processes and community context, can help to identify areas where the needs of particular groups are being met and where more attention is needed, facilitating more effective programs for all.

- Testing and evaluation should serve the needs of a multiplicity of users. Teachers may need test and evaluation information to make

instructional decisions; and local school and district administrators, as well as policymakers at the state and federal levels, need such information to guide their planning and decisionmaking. If they are to be useful in supporting and improving schools, evaluation and testing activities should be decentralized to the local level, while at the same time maintaining their utility for addressing legitimate public policy concerns at state levels in particular.

- Testing, evaluation, and standard setting are endeavors which are partly technical, partly political, and partly social. Technical expertise is essential in test development and analysis, to ensure the valid and reliable use of test results; social understanding is essential to ensure fairness and utility. Similarly, evaluation questions arise out of people's information requirements, while the design and interpretation of evaluations depend on technical competence. The definition of standards depends on values and consensus; the measurement of their attainment involves technical considerations.

While we are optimistic about the potential of educational testing and evaluation, we also are aware of their current shortcomings, cognizant of their potential misuses, and sensitive to their possible unintended effects. A national center must play a vigilant role with regard to these concerns and functions as a consumer advocate to the field, analyzing current practices and informing public policy.

Problems in Practice

Research in educational testing and evaluation has made important strides in the last decade and its methodologies hold great promise for improving the state of education. Nonetheless, significant problems remain in educational practice, problems related to the quality and diversity of existing measures, to the validity of the inferences that can be derived from these measures, and problems related to their utility to and impact on the educational system:

Problems related to quality of information.

1. Most of the testing and evaluation procedures currently used to assess students, programs and schools cover only a narrow range of the knowledge and skills that are the targets of schooling and do so without adequate attention to the nature of these knowledges and skills. For example:

- The National Council of Teachers of English have long decried reliance on multiple choice tests as measures of writing skills. Associations of teachers of mathematics, of social studies, and of science have similarly criticized the content of existing tests and the levels of achievement which are assessed.

- In the push to implement new testing programs, some states and school districts have paid more attention to new psychometric

techniques than to the knowledge domain being assessed and its cognitive underpinnings.

2. Given what is known about testing and evaluation design, tests tend to be of poor quality. For example:

- o The testing materials most commonly used by teachers, e.g., end-of-chapter tests, are often extraordinarily poor. They can mislead the teacher into believing that students have learned when, in fact, they have not; or that remedial exercises are needed when, in fact, more advanced materials would help to enhance learning.

- o The bells and whistles of the computer revolution and its slick print-outs often give an undeserved aura of scientific rigor to score reports. What the reports fail to convey is the arbitrariness of many classifications (e.g., "mastered" vs. "failed to master") and the poor reliability of the information, which may be based on only two or three items per skill.

3. Bias in the assessment of achievement for special groups is a continuing problem. For example:

- o While concerns for bias have alleviated many problems of stereotyping, teachers report that many formal tests are unfair for their students.

- o Sophisticated psychometric techniques have been developed to identify biased items but the source of the identified bias often remains unknown.

4. The quality of measures at the post-secondary level is particularly problematic. For example:

- o College admission measures serve as the primary indicator of the entire precollegiate system, ignoring other important outcomes and alternate postsecondary experiences. These measures, in addition, are not well articulated with either precollegiate curriculum or with post-secondary course offerings.

Problems related to quality of inferences.

5. Most testing programs and evaluation systems devote scant attention to the mediating factors, e.g., the quality of educational processes, background variables, and other contextual characteristics, which are basic to understanding student performance. For example:

- o Every year, a metropolitan newspaper in California ranks schools in terms of their students' scores on achievement tests. Missing from these public reports is any consideration of the factors that may explain differences or changes in rank, such as a sudden influx of children from different language backgrounds, high transiency rates, and absence rates.

6. The Federal concern for developing a National Report Card underscores the need for state and national level indicators of overall educational quality, but many problems remain. For example:

- o The component indicators of quality receive considerable attention but tend to focus on grossly, uncertainly defined but more easily accessed datasets of macro variables, e.g., dropout, student "achievement" data (like the SAT examination), teacher academic history. Neglected is the broad picture of input, process, and outcome indicators which might provide the critical context for understanding and judging comparative quality.

- o Potential sources of valid student performance data exist in ongoing state assessment programs, for instance, but investigations of means for aggregating such information are only just underway for state by state comparisons. The importance of test content receives less attention.

7. Concern for student achievement and the quality of American education escalates each time an international comparison of student performance is conducted. Yet there has been little consideration of the use of international studies, or the measures generated by them, as benchmarks to protect America's ability to compete in technological, academic, and economic futures. For example:

- o The Second International Mathematic Study provided a comparison of the United States and 20 other countries. Results show that the United States performed relatively poorly in comparison with Japan. Less serious consideration was given to the meaning of these data with respect to the role that content coverage, the quality of instruction, or the differences in background, abilities, and attitudes might play in the highlighted performance differences, although data are available on these student and instructional characteristics are available.

8. Because different types of decisions (e.g., policy, institutional, instructional, counseling) require different types of information, a patchwork system for collecting information has been created. Not only are the testing and evaluation procedures used unnecessarily intrusive, but the information produced is overly redundant. The redundancy may be particularly acute for special populations. For example:

- o Children participating in a Chapter I program at a midwestern school must take the CTBS in the fall and again in the spring, in addition to mandated state assessment tests, a districtwide norm-referenced test, and an array of curriculum-embedded tests. The information from these tests is never integrated is largely redundant, and only tangentially influences teaching practices.

Problems related to utility and impact.

9. Student testing programs on which much of evaluation depends, are externally imposed, from the top-down, but the use of data for local school

improvement is a bottom-up proposition, local and specific in nature. The result is data of limited utility for teachers and school administrators. For example:

- Extensive interviews with district administrators, principals, and teachers in one midwestern school district found that while each of these groups believed the tests had value for the system as a whole, each group also said the tests were not germane to its own needs. Thus, district administrators said that tests were helpful to teachers; teachers thought them useful to principals and principals felt they were essential to district administrators. In short, no group acknowledged that it found such information valuable.

- According to a national study of teachers' use of testing, teachers reported very little practical decisionmaking based on formal testing because of the mismatch of test content and instruction, poor reporting formats, and inappropriate timing of results.

10. Schools are supposed to be vehicles of social mobility and equity, giving all students an opportunity to achieve and to reap the benefits of productive participation in society. Although rigorous testing systems are supposed to contribute to this process, evidence suggests that testing may actually impede social mobility. For example:

- According to a prestigious national study of schooling, testing has contributed to the tracking of students into rigid vocational and academic lines, thereby reducing the prospects for individual growth and satisfaction.

- The treatment of special populations (e.g., children from different language backgrounds or with different developmental histories) often amounts to placement in dead-end tracks with little opportunity for change or advancement.

11. Tests and evaluation are regarded not only as processes for assessing educational quality, but as significant interventions in themselves that will promote excellence and high standards. There is widespread belief that the imposition of testing systems will focus and motivate learning, but other effects contrary to excellence may also accrue. For example:

- One eastern school district, echoing teachers' concerns in a national study, reported substantial narrowing of the curriculum, away from science, art, history and higher level skills and toward the basic skill areas assessed on mandated tests.

- Acceptable pass rates are a political necessity, resulting in cut-scores that reflect neither excellence nor even minimum competency.

These three problem clusters, quality of information, quality of inferences and interpretation, and utility and impact of testing and evaluation reforms are central in the conceptual framework underlying the proposed research program. This conceptual framework is described next.

Assessing and Improving Educational Quality: Conceptual Framework for the CSTES

We take as a point of departure a model of the Educational Quality Improvement Process (EQIP). This EQIP model portrays the role of testing and evaluation in improving educational quality. The model is grounded in our understanding of the nature of the educational context, which we explicate next. Two critical requirements for the model are then described, validity of information and quality of inferences; the effects of these requirements ultimately is judged by theirutility and impact on educational quality. These requirements and their impact are the focus of a substantial portion of our R&D program.

Our goal is to conduct R&D that contributes both to better understanding of educational quality and to its development as well. Our simplified picture of the role of testing and evaluation in improving educational quality is presented in Figure 1.

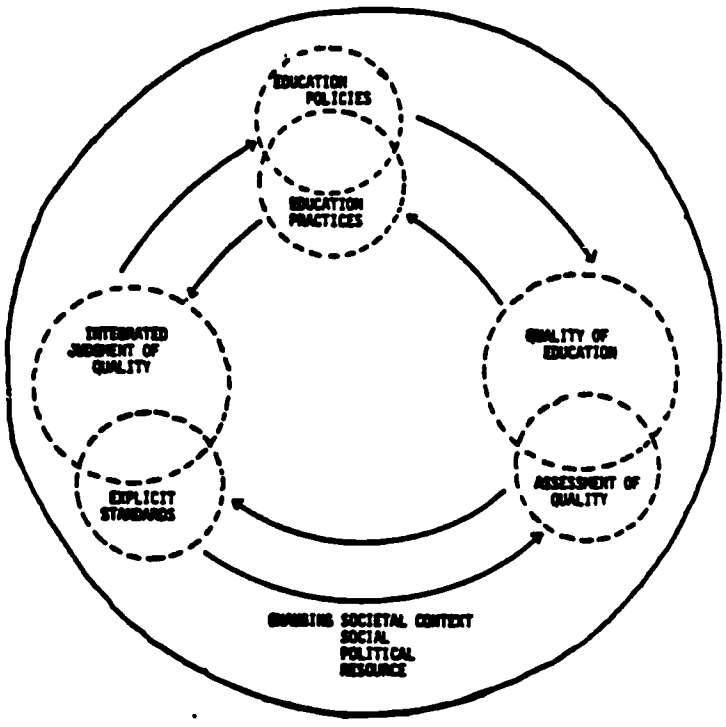


FIGURE 1: THE EDUCATIONAL QUALITY IMPROVEMENT PROCESS MODEL (EQIP)

The Educational Quality Improvement Model (EQIP)

The model displays the interaction among the formulation and implementation of educational policies and practices and the assessment and judgment of their quality. At the simplest level, educational policies are formulated to influence educational practices. But much of educational practice also develops bottom-up and on an informal basis. These existing policies and practices create the actual level of educational quality experienced by students and teachers. The next step is an assessment of educational quality, a process that can address only partially the true quality of effort and its effects. Following assessment, judgments are reached about how well policies and practices are working. These judgments may be strongly influenced by explicit standards but also develop from a wide source of other values. The model is arrayed in a circle to indicate that this process is neither discrete nor linear, and its components are set in important contexts which significantly affect and are affected by their operation. We have described one point of entry in the model, starting with the formulation of educational policy. Taken at another entry point, judgments of quality (substantiated or unsubstantiated), or attention to explicit standards, lead to assessment or assessment policies and practices which in turn affect other educational policies and practices. Here assessment is acting as an intervention. From a third entry point, assessment of quality can identify needs for new interventions in policy and practice, which are subsequently assessed, judged, and become the subject of continued or modified action.

Throughout the model, there is recognition of both implicit and explicit meanings and realities and of formal and informal sources of information. (Lindblom and Cohen (1983) have been informative on this point.) For example, the model recognizes that formal policies provide only general guidelines and exert imperfect control over actual practices at the various levels of the educational hierarchy. Second, policies and practices are dependent on formal and informal assessments which provide a narrow and imperfect estimate of reality. Third, the model recognizes that judgments about quality require the integration of various sources of information against general values and expectations for education, only some of which are represented in explicit standards. Fourth, the model acknowledges, with the intent to explore, the effect of contextual factors on the assessment and judgment of quality. These factors include changing policy expectations, social, organizational, political, and demographic factors and resources which are in constant flux and which can only be grossly approximated for any period of time.

Educational Contexts of the EQIP Model

The EQIP model is grounded in our understanding of how the educational system operates. Below we present three views which are essential to our understanding. The first is a hierarchical view of the multiple policy and administrative levels responsible for the educational system. The second is a longitudinal view of the educational system and its interdependent segments. The third is a pluralistic view of the system's clients, its students.

Hierarchical view of the educational system. Figure 2 depicts a hierarchical view of the multiple policy and administrative levels which are responsible for the quality of educational policies and practices. While the picture portrays the system as a neat configuration of nested entities, the concentricity of the circles is neither neat nor closed. A hallmark of the American educational system, and one which complicates both its evaluation and governance, is that the system is "loosely coupled" (Weick, 1976), with each of the lower levels exerting significant independence.

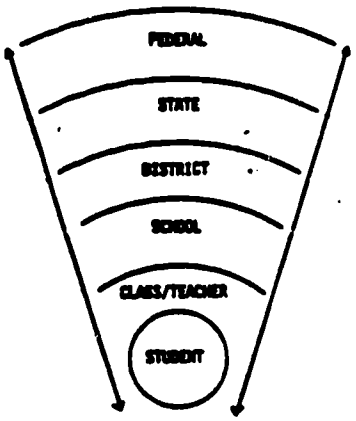


FIGURE 2: A HIERARCHICAL VIEW OF THE EDUCATIONAL SYSTEM

The figure shows the student at the center, as the primary client and ultimate recipient of educational quality, surrounded by the various contexts which influence the quality of education: classroom, school, local district, and state educational institutions as well as national, international and socio-political contexts. At the postsecondary level, this picture would omit "local district," except for certain community college venues. For private institutions, the state level may or may not have relevance. The intent of this picture is to illustrate that policies at various levels, translated into actual educational practices, have successive impact, with direction of impact both outward and inbound (that is, "bottom up" and "top-down.") These policies may have direct impact on students in the case where they completely traverse the entire system (e.g., lengthening the school day). Or the policies may affect students less directly and depend on a chain of assumptions about the relationships between certain factors and educational quality (e.g., raising teacher salaries).

The point is that policies and practices at all levels, and the interactions among them, affect the ultimate quality of education

experienced by students. To improve policies and practices, as well as to promote accountability, we believe that all practitioners and policymakers need information about educational quality (i.e., information about the quality and consequences of students' classroom experience). Overlapping assessment systems have mushroomed in an attempt to provide such information for each level (e.g., routine classroom assessment, district evaluation programs, state assessment) yet these assessment systems, like their corresponding organizational structures, are not necessarily congruent in focus.

Longitudinal view of the educational system. While the hierarchical view describes the multiple administrative levels involved in the system, the longitudinal view is concerned with multiple institutional levels. The longitudinal picture is essential to examine the quality of the system as a whole and to assess its effectiveness in educating and preparing the populace for productive lives. Figure 3 presents this longitudinal view of educational services and outputs, displaying the path a student takes from school entry through critical transition points to various exit points: entry to elementary school, the end of sixth grade, the end of junior high school, the end of high school and various post secondary options (sometimes commencing before formal graduation, including traditional college and university enrollment; technical tranining, employment, the military, and non-productive outcomes (unemployment, incarceration, etc.).

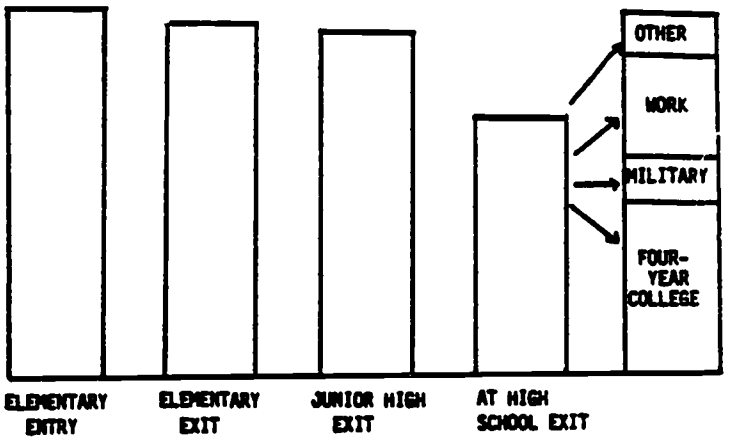


FIGURE 3: LONGITUDINAL VIEW OF EDUCATIONAL SYSTEM

What is the relationship of quality assessment to this longitudinal view? First, there are both short-term and longer term effects. In the short term, the success of students at any point in the continuum can be used to estimate the cumulative effects of earlier educational services, conditioned, of course, by contextual variables. Taking the longer term view, the figure reminds us that there are various legitimate outcomes of education and that choices other than college for students should be included within the educational quality assessment paradigm, e.g., success with business and corporate training requirements, the entry and retention of individuals in employment, and their entry and success in the military. (SAT scores and other measures of college preparation, in other words, give at best an incomplete picture of educational quality.)

A second implication of the longitudinal view is the obvious inter-relationships between educational levels and the need for articulation at the identified junctures. Clark writes eloquently of the mutual effects of precollegiate and post-secondary systems, effects that often are mediated by tests and other assessment and placement devices; his words apply to junctures between other levels as well:

> We may conceive of the relation between secondary and higher education at the outset as a two-way street along which the nature of traffic in one direction is quite different from the flow of people and activities in the other. Up the street, from the "school" to the "university," we encounter primarily a flow of students. The school selects them, trains them, orients them, certifies their competence, and sends them on... Whatever the quantity and the quality, and the degree of opportunity, the school clearly shapes the human resources made available... In education, generally, an impelling principle of sequence gives lower units this particular role in determining the nature of higher levels.[1] Down the street, from the university to the school, the traffic is different, consisting always of two major vehicles of influence. One is personnel... A second vehicle is curricular in nature: the university sets course requirements for its own students, and often itself sets entry requirements that influence what teachers will teach and what students will study in the school. Students who want to go on must master those materials and pass those examinations that permit them to be a part of the upward flow.
>
> Burton R. Clark (1985)

Just as the hierarchical view of educational systems highlights the need to be sensitive to the needs of various levels of the system hierarchy, the longitudinal view encourages sensitivity to various levels of schooling. For example, the modal organization of elementary education is the self-contained classroom, resulting in the need for multidimensional indicators aggregated within classrooms (e.g., performance in different content areas, self-concept, attendance). The departmental organization at the secondary level presents an opportunity for more content detail and the challenge of aggregation of students across teachers and blurs lines of accountability for basic skills. Explicit course choices and differing educational goals contribute additional complexities to indices of quality at this level, as do issues of different classroom organizations and of problems of articulation among grade levels.

The pluralistic view of the educational system. The dramatic diversity of students served by our schools provides a third important context for our conceptual framework. Students come from a variety of backgrounds, ethnicities, and communities. They exhibit different ability levels, cognitive approaches, language facility, and interests. Many students have special needs for educational attention: physical handicaps, learning disabilities, and or highly developed talents or aptitudes. Students aspire to the full range of accomplishments a formal

educational system can provide. And the American system is committed to helping them to satisfy such goals. It has assumed dual responsibilities for addressing individual differences while communicating the common learnings necessary for an integrated and unified society.

ecognition of the energy and richness of our student population brings a serious set of concerns to the picture of educational quality. We confront choices related to how much attention should be paid respectively to maintain diversity or to increase commonalities in our educational policies and practices. We must also take into account student differences as we attempt to assess the quality of the policy choices we make. Here our concern relates to the fairness issue. We balance equity interests with attention to standards. Different standards for minority students, for instance, increases diversity but reduces the fairness in the long run. Technically, we have advanced in methods for detecting bias in measurement, and for correcting statistical differences attributed to varying levels of student performance, but much remains to be done in the area of what and how we assess our diverse student body and how we make confident inferences from our findings.

American educational variety also grows from the diversity of the communities in which our students reside. They may live in urban, suburban, or rural settings. Their communities may be stable or radically changing. The economic productivity surrounding them may be vigorous or tenuous. Their schools may be large complex organizations or smaller, more personalized settings. They may be very like other students in their class in background or represent a minority of one or another type. Attempts to improve educational quality and to assess its quality will succeed only to the extent that these important factors are considered in our analyses and our actions.

Our EQIP model and the three contexts above present a backdrop for our approach to assessing educational quality. For the model to be successfully used, two critical requirements must be met. One requirement is valid information; the other is quality inferences derived from that information. Clearly, these requirements are necessary but not sufficient: Social, organizational, political and simple, human preferences influence our policy choices and our interpretations of their success. But valid information and high quality inferences are at the core of the EQIP model. They are amenable to conceptual study and empirical improvement and are appropriate to the fundamental issues to which the CSTES is directed.

Requirement One: Validity of Information

Validity of information includes concerns with the accuracy, representativeness, and comprehensiveness of what is claimed to be measures of educational services and effects. While it is tempting to cast this argument broadly in terms of the R&D needs of full range of information that might be included in the assessment of education, we plan, at the outset, to focus substantial effort on the issue of student achievement and performance. The reason for this decision is not to disdain the utility of other indicators of performance; to some extent we will address these as well. But rather we believe that this R&D Center

should help others to estimate teacher quality, school effects at the elementary, secondary, and postsecondary levels, impact of state and local policy, and content-based work in the disciplines rather than devoting the majority of our resources to areas to be addressed by other institutions. Nowhere, however, is the responsibility for exploring fundamental issues in student performance more properly assigned than to the Center for Student Testing, Evaluation, and Standards.

We believe that much of the present information on student achievement is based upon anachronistic models of human learning and often does not reflect the best availablepsychometric and statistical models. Improving the quality of student achievement information depends on relatively sophisticated notions of validity. Four points deserve particular attention: content quality; appropriate approaches to content assessment; cognitive basis of individual differences; and assessment purpose.

Content quality. Research on student assessment must take into account the content of what is being tested. General levels of content specification must be augmented to reflect research on cognitive knowledge representation as well as to the more significant concepts in the field, as judged by scholars in the academic disciplines (Haertel and Calfee, 1983). Quality content must be at the core of any measures. With respect to subject matter content, recent advances in cognitive science are pertinent (Larkin, et al, 1980). They suggest that models for content-based analyses must be specific to subject matter structure for the design of procedures to find out what students really know and can do (Shavelson, 1983). These procedures may be much more susceptible to differences in the way content is organized within the discipline. If so, then much less general rules of thumb for achievement for achievement test design will be required, and one challenge will be exploring the limits of general test development procedures versus the need to create separate models useful for assessing different content areas.

Appropriate approaches to content assessment. Validity in measurement also depends upon the belief that the means available for assessment are appropriate to the subject matter. A case in point was the dependence upon multiple choice measures to assess students' ability to compose essays. Though logically indefensible, this practice persisted because of the ease of computing reliability estimates, the low cost of data collection and scoring, and the reliance on correlations to show that ability in composition correlated with these measures at some respectable level. Research studies of writing assessment (Hays, et al, 1980) however, demonstrated that the cognitive demands of written composition were vastly different than those of selecting responses. The development of practical, efficient, and reliable scoring strategies combined with these cognitive analyses to permit the more valid assessment of this critical skill area (Quellmalz, 1985). Similarly, it may be demonstrated that certain problem solving tasks in science or analytical tasks in comprehension of literature may be better assessed by means other than traditional multiple choice or short answer tests. Determining what options there are and marrying those findings with what scholars feel are sensible approaches to the assessment of their content areas could result in a broader, differentiated mode of student achievement measurement.

Cognitive bases of individual differences. A third issue in valid achievement assessment is the extent to which any approach permits students to demonstrate the most, or the best that they know. Bringing back the attention of researchers to the individual differences among students test-taking preferences may allow us to assess more accurately what educational effects are. Attention to alternative symbolic representations of task (aural, pictorial and dynamic) and to options in response modes, as developed from cognitive and subject matter analyses, may allow the creation of more diverse testing systems. These options can help to overcome the criticism of uniformity, triviality, and narrowness of current testing practices and reflect more directly the reality of the enormous variation in cultural, experiential, and learning histories of our students. What we need to explore are alternative options for teachers and students to demonstrate educational achievement, options that are not easier or harder, or preferred rather than undesirable, but assessment choices that share rigor and credibility. If this exploration is successful, our contribution to the validity of information will be clear, and ideas about what "difficulty" of tests means might undergo redefinition.

Assessment purpose. A fourth area of validity in achievement measures relates to the purpose for which the measure is used. While achievement test purposes are commonly thought of as diagnostic, placement, monitoring and certification, with different models of testing proposed for each, our particular interest focuses on validity as it relates to student learning in the instructional context. An important issue is the extent to which a single assessment system is valid for a variety of purposes. The types of tests which teachers most frequently use and accept (Herman & Dorr-Bremme, 1983, Dorr-Bremme, 1983) deserve particular attention. Such systems need to combine attention to design, psychometrics, and new technologies. Research will explore ways to increase both the validity and utility of such systems.

Requirement Two: Quality of Inferences

The central thrust of efforts to improve the quality of inferences from educational information is to build our confidence that the bases for judgment, evaluation, subsequent action, and consequent impact on the educational system are as accurate and circumspect as we can make them within existing knowledge and resource constraints. The issues here are legion. First, we have concern with the proper linkage of information to any given primary decision context. Second, we are concerned for the multi-level, multi-institutional use of information. What distortions occur when information collected for one purpose is applied at another level? Third, we are interested in economy, to avoid burdening the system with more and more information of less and less utility. Methodological options for creating linked data bases may provide a solution. Fourth, we are interested in the comparison issue. Given a set of information, how do we know what to make of it? Last, we maintain an interest in expanding, as appropriate, the information base to ground and elaborate our interpretations.

Multi-level inference. Let us treat the first three questions together, as they pertain specifically to the multi-level, multi-institutional problem, remembering that inferencing is at the core of evaluation processes. A primary problem is that evaluation practices at each level of the educational hierarchy operate relatively independently from one another. State level information is specially designed, collected, and analyzed for state purposes; districts use a different set of measures for their decision needs; to the extent that teachers use formal information sources for their instructional decisions, they tend to rely on those provided with curriculum materials or developed on an ad hoc basis. On the surface it seems reasonable to assume that different measures are needed to meet the unique decision context at each level, and it might be argued that an overlapping testing strategy permits triangulation that supports validity. However, there are serious problems within such an approach, with tensions between the need for more generalized measures as one moves up the hierarchy and the need for sensitivity for what actually transpires at the lower levels. For example, a primary function of achievement testing at the state level frequently is to ascertain what students are learning with regard to a state curriculum framework. The framework is typically specified at a general level as are the measures which assess it. The resulting assessment may or may not be sensitive to either the variations in specific curriculum implemented at the district level, or to variations in instructional programs implemented by teachers in each district. Or more to the point, we can be sure that at least some mismatches will occur at each level, mismatches that compromise the validity of the assessment for some purposes. The assessment will always miss its mark and add both noise and valid information to the system.

Other problems arise when there is no common basis for inferences about educational policies and practices at the various levels. The general intent of educational policy formation is to improve the quality of educational services and to help our students attain the highest levels of competency in school subjects. At some time, the policies need to be translated into practices that are compatible with understandings at the levels of real implementation -- ultimately with what teachers and students see as their requirements and day-to-day practices. There is high potential for slippage when the information used to assess quality and formulate policy functions independently from that used to actually teach children. While it would be neither appropriate nor profitable to envision a fully articulated system where information useful for instructional decisionmaking is also employed at the highest policy levels, the present low level of overlap creates special and persisting anomalies. It also causes unnecessary costs -- in financial resources devoted to test administration and scoring and in opportunity costs related to teachers' and students' instructional time. Current duplicative systems, in other words, may be both ineffective and uneconomical.

Part of the effort of this Center will be to explore the limits of common or compatible information bases for multilevel educational decision-making, particularly in the area of student achievement. In the name of economy, of preservation of student time, and of quality

inferences, we believe attempting to move the various levels toward some larger proportion of shared data has merit. Embedded in this problem are methodological issues related to integrating horizontally different kinds of information appropriate within a level, to linking and equating locally sensitive measures, and to summarizing and integrating information vertically (or decomposing down) so that the appropriate level of detail is available for information users (Baker, 1983; Burstein, 1980, 1981, 1983). All these methodological issues again are nested in educational contexts that must be taken into account substantively and methodologically to reflect the special character of different levels, and facts of individual differences: among children, teachers, schools, communities, districts, and states.

We are also interested in the extent to which common or linked quality assessment can inform us about the cumulative effects of education across the longitudinal view of the system, as presented in Figure 3. Here the concern is to include indicators that are sensitive longitudinally to educational quality as exemplified at different institutional levels, e.g., elementary, secondary schools. At present our information is woefully limited. Can we tell if student effort is qualitatively maintained, increased or decreased at identified institutional points. Can we estimate cumulative effects? Can we assess the articulation of programs across school levels? In general, the answer to these questions is a resounding no. Our interest, then, is to develop measures that have clarity and continuity. And we need ways to link information between grades within particular institutions and across institutions to provide ecologically valid inferences about student progress over time.

The potential benefit of a multi-level, multi-institutional approach to interpretation is clear. Not only could the intrusiveness of testing and measurement be reduced, but the validity and linkage of inferences could be enormously strengthened when policymakers and teachers share a common core of information (if not at the same level of detail) to guide both their policy formation and educational practices.

Comparative inference. As noted earlier, valid inferencing raises questions of comparison. Despite wishes and dreams to the contrary, comparison is an important fact of life in educational evaluation and policy assessment. Although the habit of comparing students on percentiles has waned as the favored metric of educational quality, there is strong and abiding concern with the comparative quality of educational services, organized in schools, districts and in states. Comparison is at issue in determining the merits of regular, on-going educational enterprises, but is more readily understood in the context of judging the effects of an intervention.

A first approach for judging the cumulative impact of an intervention is its effects over time on existing indicators regularly used to track educational practice. These may be as homely as regularly administered standardized tests with all their known technical limitations but undeniable public credibility. Or a broader range of indicators could include dropout rates, attendance, and performance on tests sponsored by administrative levels beyond the school (such as district wide competency

measures, or state assessment.) Using trends over time for comparison is a complex matter because of changes in measures, trend interpretation, cohort differences and the operational meaning given to measures in different localities, and so on. But looking over time to determine whether student performance and regularly tracked processes improve on a range of indicators is an obvious and important first step in assessing progress.

A second kind of comparison fits within more traditional concepts of "external criteria" where broad effects are gauged inferentially by analyzing indicators remote from the school sites where education takes place. A principal example, is the use of postscondary indicators to judge the quality of precollegiate education. Witness the enormous attention paid to the decline in SAT scores which are interpreted as evidence of the decline in overall quality of schools. This approach has a number of problems. Even putting aside contention about the meaning of such blended aptitude and achievement measures, postsecondary admission statistics can no longer alone serve as unquestioned standards for precollegiate educational effects. For one thing, a singular focus on college admission misses the goals and organization of the comprehensive high school and the diversity of its student goals (Sykes, 1985). But even for the population segment aspiring to postsecondary education, the use of admission performance and acceptance rates is not easily interpreted because of contextual circumstances or conditions. For example, the pressures on postsecondary institutions to fill available student slots, coupled with the traditional committment in the United States to open access to postsecondary schooling, make college intake numbers less convincing as indicators of public school performance. What might be credible measures suitable for comparison are what happens to students in college, how they perform, how they demonstrate the quality of their academic preparation (Pace, 1983). What sense should be made, for instance, of the extensive remedial efforts now required by two year colleges and even prestigious research universities for their entering students? Certainly these efforts suggest that the quality of schooling cannot be easily glossed over in terms of distributions of students moving to higher education. A serious effort in this arena opens up the questions of what postsecondary education is, who it serves, and what its effects could be. Clearly, postsecondary institutions have conducted evaluation efforts, directed at ranking on institutional criteria faculty, libraries, research productivity, and so forth. But for postsecondary information to serve more than a mystic, habitual indicator of public school preparation, the quality of student learning in college will need to be directly addressed and soon.

Another obvious comparison option is the relative quality of schools (districts, states) with respect to a national standard. Because of the local organization of education, a clear criterion for comparison has not existed. But there remains a continued tension between the desire for a "national" picture and the local authority for educational services. The pull of a national achievement indicator is attractive, but resistance is also strong for constitutional and for less lofty reasons. A national test could be created (and is periodically suggested), but only at the risk of reduced local validity of findings for diverse student and instructional settings. The tradeoffs of uniformity vs. some direct measure of national

performance have been partially addressed by the National Assessment of Educational Progress. But since its design was originally not intended to provide comparisons linked directly to the educational efforts of bureaucratic units (Wirtz and LaPointe, 1977), necessary changes in the frequency, types, and distribution of NAEP test administration could sharpen contention and reduce compliance. Alternative processes for providing more valid national comparisons are under development (Burstein and Baker, 1985; Bock, 1985) related to the use of existing state level indicator data to feed into analyses conducted by the National Center for Educational Statistics (Elliott and Hall, 1985). But unless there are significant policy changes, the quest for a national comparative base for educational impact will continue to be satisfied by partial, qualified, and in some minds, appropriately blurry information.

A last arena for comparison that has grown in attention is American educational quality contrasted to that produced by other countries. Through international studies such as those conducted by IEA (Purves, 1980, Travers, 1984, Burstein et al, in press; Baker, 1985), the standing of US students is judged on internationally arbitered performance measures. The utility of inferences from international comparisons can easily be challenged: educational systems differ dramatically in terms of tradition, size, centralized management, tracking, selection, and access of students. On the other hand, such comparisons do provide an imprecise but compelling benchmark: when all things are considered, how well do US students do? Yet, any international comparison should also answer the question of what else can US students do and where else do they show deficiencies. At any rate, it is dangerous to assume that education in the United States should adopt Japanese instructional practices or French or British examinations systems. Clark (1985) points out countries which emphasize high school exit rather than college entrance examinations have traditions of academic excellence, prestige for teachers teaching the highest track. These countries can demonstrate tight linkages between secondary and higher education excellence when there is concommitant tight tracking and selection processes in the lower schools. Many of these conditions run smack into American traditions of access and equity, the historical, if unfortunate role of teacher education in the University, among a complex of factors. So inferences from such international comparisons may create general competitive goals rather than a specific all to adopt practices of other countries. These inferences should be made carefully, and must attend to systemic differences in the organization of education as well as the surface features of examination processes.

Expanding the band of information. It is a fact that much of precollegiate school evaluation activity depends upon measures of student achievement. The usefulness of such information depends upon not only validity issues identified earlier, but on the extent to which such information adequately represents educational quality. It is our judgment that the present dependence upon achievement tests grossly underrepresents important dimensions of educational quality. Just as we hope to expand the base of valid measurement of achievement, we also wish to expand the range of information used in evaluation systems beyond achievement, to include other important indicators of quality (Sirotnik et al, 1983). Construct validity in an achievement area has been pursued by combining various

achievement tests. Here we are pursuing the combination of achievement measures with other indicators to assess more validly a larger construct, that of educational quality.

To obtain a full picture of educational quality, properly contextualized, is probably a fool's errand. To obtain an improved picture, with broader focus, is within our grasp. Our measures should address variables of context, inputs, processes, and outcomes. Context measures include characteristics of children, including language proficiency, socioeconomic conditions of settings, transiency rates, and so on. These facts often are overlooked in simple evaluation studies and often seriously influence appropriate inferences about educational quality. Input variables include measures of financial support, quality of teachers attracted to the system, quality of physical surroundings, etc. Process variables involve the interactive behavior of administrators, teachers, students and parents, students' instructional activities, including such things as time on task, expectations for learning, parent involvement, and teacher satisfaction. Outcome measures include standardized achievement tests, measures of student production (such as writing), student attitudes toward school and learning, dropout etc. Here the concern is selecting variables that are likely to be relevant to the intervention assessed and selecting measures that meet criteria of validity similar in scope (but not in nature) to those identified for achievement measures in the section above. In selecting variables and measures, our interest is in identifying an optimal number for sensible interpretation. Of special emphasis is the relationship among process and outcome measures, especially the extent to which changes in process may serve as proximal predictors of student outcomes. This particular concern derives from the checkered history of comparative evaluations where measures of instructional process were rarely undertaken, or when processes were measured, treatment differences were often undetectable (Burstein, 1981; Stake, 1978; House, Glass, McLean, and Walker, 1978). Measures of organization process (Williams and Bank, 1984) also appear to be important predictors of intervention effects. We do not see the mission of the CSTES as being principally concerned with the identification of these variables, for this task is better accomplished by other R & D centers (related to specific levels of schooling, teaching, etc.) Furthermore, expansion of the set of educational quality indicators, although a strong interest in our present proposal, is also being addressed by other organizations, such as the NCES and National Academy of Science, to name but two of the main actors. Our interests are to assure that places are held in evaluation systems for such variables and to assist in the measurement issues attendant to their application.

How shall indicators be conceived? The economy and validity demanded by multi-level application of measures should be a concern as we attempt to broaden our information base. Clearly, the prescience that information will be applied at different levels will influence the nature and form of the questions posed. If economy of effort is a serious matter, then agreements must be made on apparently simple matters such as format and meaning of variables. These agreements about the range, type, real meaning, and formats of information will generate tension in the vertical operation of the system (among information providers and users from the

classroom, school, local district, and state levels). Different but equally important issues will need resolution as coordinate information needs generalize horizontally (State to State, for instance). And another set of contentions will be addressed by members of different institutions, e.g., elementary and secondary schools, community colleges, and universities, who attempt to find indicators to assess the articulation and cumulative effects of multi-institutional systems.

Integrated educational quality assessment: Creating multi-level evaluation systems. Integrating quality information, valid inferences, and multi-level and multi-institutional contexts into a set of operating systems is a tall order. To recap the discussion thus far, features of such a ideal system would consist of valid information including student achievement measures (using a variety of methods and formats), and an expanded set of indicators of school context, processes, resources, and non-achievement outcomes. Functionally, valid inferences meant be drawn by integrating measures into valid composit indicators, interpreting information in the light of the specific and multi-level context(s). Last, the system would provide comparisions against multiple criteria. The intent of this system would be to generate ways to evaluate educational policies and practices, and would contribute to their amendment and improvement. Clearly the nature of the educational system precludes a lock step development of even an approximation of such an evaluation system. It would certainly be naive to expect, for instance, that the imposition of a particular set of state level standards of testing and evaluation requirements would have uniform or generally consistent effects as the intent of policy was successively reinterpreted at lower levels of educational organization.

The abstraction of a complicated system takes on unexpected forms of reality as real implementation is addressed. Our intentions in exploring the design of multilevel systems involve a dual focus on the technical quality of the intervention and on the local realities that contribute or impede the implementation of innovation (Hathaway, 1985; Cooley and Bickel, 1985; Sirotnik and Burstein, 1985; Bank and Williams, 1984a; Herman, 1985; Dorr-Bremme, 1983). When grand plans confront habits of daily decision-making in classrooms and schools, grand plans often crumble. Thus, in our own efforts we intend to provide opportunity for local participants, including teachers, school principals as well as district and state managers to have serious influence on the shape of these evaluation systems. We hope to balance, in fact, the locus of evaluation systems at the local level (bottom-up) with the clear requirements of state and national policy. We also intend to conduct intensive studies of implementation so that our efforts may be successively adapted to work to the satisfaction of the research scholars, the policymakers, and the people who conduct the day to day business of teaching and learning. We would expect such preliminary systems to incorporate the best R & D available, from whatever source, in their systems. We would expect these systems to function in a formative or improvement-oriented manner. Should the systems have merit, we would then wish to assess their impact as interventions affecting educational quality.

Utility and Impact

The foregoing discussion of the EQIP model, its contexts and its requirements for quality information and inferences is incomplete. The power of our formulation also must be judged in light of its utility and impact in the real world of schools. Successful application of procedures derived from such a framework will depend on less technical concerns. One of these is the utility of the information generated by testing and evaluation processes.

Utility. Utility can be analyzed in at least two ways: perceived utility and objective utility. Perceived utility resides in the eye of the user. Information can be thought to be useful, described as influential in ways of thinking about problems or in actual decisionmaking. Information may provide clear guidance related to a particular purpose or shed light in an unexpected way on an unresolved issue. In this area, we depend upon reports of individuals regarding usefulness, or infer utility from the ideas held, language used to express ideas or actions related to extant information (Glass, 1972; Weiss, 1977).

Objective utility involves the analysis of consequences of information for decisionmaking. In some sense it is a reverse engineering problem, a problem of tracking back from decisions and attributing partial causes to related information. This process is laborious and uncertain in the light of the weak links in chains of decisions and because rationalization of decisions is a part of organizations and policymakers everywhere. This analysis process also provides a distorted view of the ways information likely affects decisionmaking, not at all as systematically and neatly as in an experimental research paradigm with clear treatments, periods of implementation, and crystalline findings. Rather, quality information gets used irregularly, in combination with informal sources and beliefs and on a lurch and languish schedule. Research related to evaluation and knowledge utilization (Weiss, 1972, 1977; Pelz, 1985; Alkin et al, 1985) is pertinent here. We also assess the utility of information in terms of its conformance (construct validity) to findings in related areas, the extent to which information confirms trends from other data sources or can be thought to illuminate new courses of action

Impact. Objective utility then links the available information base, the inferences drawn from it to a set of decisions. Another, tougher question involves the result of the decision. What is its impact? Baldly put, did the formulation and application of testing and evaluation have impact? We have all learned, living with pollution, asbestos, food additives, and so on, that outcomes can be both positive and negative, that planned good can turn into evil. So our study of impact is goal-free (Scriven, 1974) and deals with both benefit and loss. We do not see the study of the impact of testing and evaluation to be an interesting sidelight. Nor, we are quick to say, can we imagine such studies to be anything close to direct tests of the concepts of quality information and valid inferences. But it is responsible to close the loop. We must not stop with analyses, with research ideas that contribute only to the generation of other research ideas. We must use the noisy and imprecise

information available from targeted impact studies of policy interventions as a basis for reexamining our views, our research plans, and our intended accomplishments. Because educational testing and evaluation have their power as applications in practice, we must describe and report their effects as they occur. All of this effort, however, is undertaken with no small measure of modesty. Research-based knowledge has a strong contribution to make, but is nowhere near sufficient. Our programs will succeed if they strengthen the knowledge base underlying practical day-to-day decisions. In the longer term, the spread of technology may make this utilization problem more tractable and the predicted effects more optimistic.

How can concerns for utility and impact be considered within an R&D program? They require intensive and multifacted study of the effects of testing and evaluation. We plan, therefore, to devote attention to testing and evaluation not only as ways to assess the system, as interventions themselves intended to raise standards and to improve educational policies and practices.

Summary

Our conceptual framework addresses the issue of educational quality assessment within the complex contexts of American education and provides the backdrop for the CSTES research and development program. CSTES staff is committed to explore the use of testing and evaluation to improve educational policies and practices at all levels of the educational system. Second, we are interested in testing and evaluation (assessment) methods which incorporate implicit and explicit expectations for education and which provide a more complete and accurate picture of educational quality. Third, we are interested in integrated judgments of educational quality, integrations made horizontally across various dimensions of educational quality, vertically, both up and down across levels of the educational system, and longitudinally, across institutions serving different ages of the population. Fourth, we are concerned with the usefulness and use of assessment in support of improved educational policy and practices.

Goals for CSTES

This orientation to educational quality assessment and improvement leads directly to the explication of CSTES goals. Inherent in our conceptual framework are the two institutional goals to which our work will be directed.

TO CONTRIBUTE TO THEORY AND PRACTICE UNDERLYING THE ASSESSMENT OF EDUCATIONAL QUALITY; and

TO CONTRIBUTE TO THE IMPROVEMENT OF EDUCATIONAL QUALITY ITSELF.

To accomplish these goals, we will focus particularly on five major objectives. The first three are derived directly from our conceptual framework the final two support critical R&D strategies:

1. **TO IMPROVE THE VALIDITY OF STUDENT PERFORMANCE MEASURES BY:**

   - o improving the content base of measures;
   - o improving the usefulness of measures in instructional settings;
   - o broadening approacnes to assessing student performance to increase their fairness and utility;
   - o integrating research in human cognitive processing and assessment; and
   - o exploring the applications of technology for test development, administration, and analysis.

2. **TO IMPROVE THE VALIDITY OF INFERENCES ABOUT EDUCATIONAL QUALITY BY:**

   - o developing methods for articulating information vertically in institutional and organizational contexts;
   - o expanding the band of indicators beyond traditional measures of student performance;
   - o integrating a variety of measures to provide a better picture of educational quality of precollegiate and postsecondary educational services and outcomes;
   - o conducting analyses of the conceptual and theoretical underpinnings of the evaluation process; and
   - o exploring the organizational and technical requirements for multilevel evaluation systems.

3. **TO EVALUATE THE IMPACT OF STATE AND LOCAL POLICY REFORM IN AREAS OF TESTING AND EVALUATION ON EDUCATIONAL QUALITY BY:**

   - o tracking international, national, state and local policy reforms in testing and evaluation for precollegiate and postsecondary educational systems;
   - o analyzing problems, promising claims, and effects and regularly reporting these to policy, practitioner, parent, and community constituencies; studying the effects of particular testing and evaluation policies on educational standards, quality of school life, and public perceptions to determine if such reforms have their intended results; and
   - o analyzing, in particular, the effects of testing and evaluation on populations with special needs.

4. **TO DISSEMINATE THE RESULTS OF OUR R&D TO A WIDE RANGE OF AUDIENCES AND TO HELP FACILITATE THEIR IMPACT ON THE FIELD BY:**

- o collaborating closely with stakeholders in testing and evaluation utilization and with the R&D community throughout the entire R&D process; and

- o disseminating vigorously the results of our research through a variety of media and through a wider network of researchers, practitioners and policymakers.

5. **TO SET THE RESEARCH AGENDA FOR THE FIELD OF EDUCATIONAL TESTING AND EVALUATION AND ASSURE IT WILL CONTRIBUTE TO NATIONAL EDUCATIONAL PRACTICE.**

CHAPTER THREE: Futures Paper

A Center for Student Testing, Evaluation and Standards: Assessing and Improving Educational Quality

The proposed NIE Center on Student Testing, Evaluation, and Standards will conduct research designed to improve the quality of testing and evaluation practices, seeking to increase their contribution to educational excellence and equity, their impact on local school improvement, and their role in enlightened policy making. Central to our approach is the belief that evaluation and testing can contribute significantly to educational quality and to planning and decision-making at all levels of the educational enterprise: from the individual student through the classroom, school, district, state, and federal levels. If they are to have such an impact, however, testing and evaluation must be sensitive to the complexities and realities of the schooling process, to the local and regional character of education, and to the multiplicity of constituencies who have a stake in education and its evaluation.

The CSTES represents a unique collaborative effort to advance theory and practice in the mission area. A creative national organizational structure is proposed which brings together leading researchers from the UCLA Center for the Study of Evaluation, from the University of Illinois, from the University of Colorado, from the National Opinion Research Center at the University of Chicago, and from Educational Testing Service to work on pressing educational problems. The utility and impact of the research program will benefit not only from the multidisciplinary perspectives of this prestigious group but also from the active collaboration of prominent practitioners and policymakers from across the country at all levels of the educational system -- school, district, state, and national. These collaborative arrangements will help to assure a targeted R&D program which contributes significantly to both knowledge production and to knowledge utilization.

Guiding Premises

Collaborators in the CSTES proposal have well-established credentials in the mission area and extensive experience in working together. The research agenda we proposed is guided by our shared belief in the importance of testing and evaluation in improving schools and in informing sound public policy. A number of premises are central to our approach:

- We believe that testing, evaluation, and standard setting can contribute to improving the quality of education. Tests -- when they are well conceived, constructed, administered, and analyzed -- can provide valuable insights into how individuals and classes of students are learning; they can help guide teaching, administration, and policymaking within our educational institutions. Evaluations of programs -- especially when they are seen as improvement oriented, locally useful, and iterative -- can help to guide the reallocation of resources, the modification and

improvement of activities, and the retraining of personnel. Standards -- set with due attention both to what is desirable and to what is feasible at the state and local levels -- can help to focus attention and promote accountability for educational improvement.

- We believe that testing and evaluation are important tools for promoting educational equity. Tests, when they are sensitive to individual differences and preferences in learning styles, provide a powerful means for diagnosing students' unique needs and providing effective instruction for all students. Furthermore, tests, when they match classroom instruction, can provide fair and equitable measures of student progress, measures which focus on learning accomplishments rather than background characteristics. Achievement measures as well as measures of educational processes and community context, can help to identify areas where the needs of particular groups are being met and where more attention is needed, facilitating more effective programs for all.

- We believe that testing and evaluation should serve the needs of a multiplicity of users. Teachers may need test and evaluation information to make instructional decisions; and local school and district administrators, as well as policymakers at the state and federal levels, need such information to guide their planning and decisionmaking. If they are to be useful in supporting and improving schools, evaluation and testing activities should be decentralized to the local level, while at the same time maintaining their utility for addressing legitimate public policy concerns at state levels in particular.

- We believe that testing, evaluation, and standard setting are endeavors which are partly technical, partly political, and partly social. Technical expertise is essential in test development and analysis, to ensure the valid and reliable use of test results; social understanding is essential to ensure fairness and utility. Similarly, evaluation questions arise out of people's information requirements, while the design and interpretation of evaluations depend on technical competence. The definition of standards depends on values and consensus; the measurement of their attainment involves technical considerations.

While we are optimistic about the potential of educational testing and evaluation, we also are aware of their current shortcomings, cognizant of their potential misuses, and sensitive to their possible unintended effects. We believe that a national center must play a vigilant role with regard to these concerns and functions as a consumer advocate to the field, analyzing current practices and informing public policy.

Problems in Practice

Research in educational testing and evaluation has made important strides in the last decade and its methodologies hold great promise for improving the state of education. Nonetheless, significant problems remain

in educational practice, problems related to the quality and diversity of existing measures, to the validity of the inferences that can be derived from these measures, and problems related to their utility to and impact on the educational system. The following examples illustrate the variety of existing problems within each of these interconnected areas.

Problems related to quality of information.

1. Most of the testing and evaluation procedures currently used to assess students, programs and schools cover only a narrow range of the knowledge and skills that are the targets of schooling and do so without adequate attention to the nature of these knowledges and skills. For example:

- The National Council of Teachers of English have long decried reliance on multiple choice tests as measures of writing skills. Associations of teachers of mathematics, of social studies, and of science have similarly criticized the content of existing tests and the levels of achievement which are assessed.

- In the push to implement new testing programs, some states and school districts have paid more attention to new psychometric techniques than to the knowledge domain being assessed and its cognitive underpinnings.

2. Given what is known about testing and evaluation design, tests tend to be of poor quality. For example:

- The testing materials most commonly used by teachers, e.g., end-of-chapter tests, are often extraordinarily poor. They can mislead the teacher into believing that students have learned when, in fact, they have not; or that remedial exercises are needed when, in fact, more advanced materials would help to enhance learning.

- The bells and whistles of the computer revolution and its slick print-outs often give an undeserved aura of scientific rigor to score reports. What the reports fail to convey is the arbitrariness of many classifications (e.g., "mastered" vs. "failed to master") and the poor reliability of the information, which may be based on only two or three items per skill.

3. Bias in the assessment of achievement for special groups is a continuing problem. For example:

- While concerns for bias have alleviated many problems of stereotyping, teachers report that many formal tests are unfair for their students.

- Sophisticated psychometric techniques have been developed to identify biased items but the source of the identified bias often remains unknown.

4. The quality of measures at the post-secondary level is particularly problematic. For example:

- o College admission measures serve as the primary indicator of the entire precollegiate system, ignoring other important outcomes and alternate postsecondary experiences. These measures, in addition, are not well articulated with either precollegiate curriculum or with post-secondary course offerings.

- o Testing has made its entrance in the collegiate environment in narrow enclaves: dealing with "underprepared," often minority students, in courses designed to ready students for college level work; less frequently in qualifying exit examinations related to writing or mathematics performance. But the larger question of the effects of higher education on intellectual growth and on preparation are inferred from patterns of course enrollment and grade point averages.

Problems related to quality of inferences.

5. Most testing programs and evaluation systems devote scant attention to the mediating factors, e.g., the quality of educational processes, background variables, and other contextual characteristics, which are basic to understanding student performance. For example:

- o Every year, a metropolitan newspaper in California ranks schools in terms of their students' scores on achievement tests. Missing from these public reports is any consideration of the factors that may explain differences or changes in rank, such as a sudden influx of children from different language backgrounds.

- o High student mobility rates may obscure a given school's quality of effort. Thus, in large urban school districts, only 40 percent of the children who enter a particular school in the fall will still be attending that school in June, and absence rates may run as high as 50 percent every day. But public evaluation documents almost never mention these factors.

6. The Federal concern for developing a National Report Card underscores the need for state and national level indicators of overall educational quality, but many problems remain. For example:

- o The component indicators of quality receive considerable attention but tend to focus on grossly, uncertainly defined but more easily accessed datasets of macro variables, e.g., dropout, student "achievement" data (like the SAT examination), teacher academic history. Neglected is the broad picture of input, process, and outcome indicators which might provide the critical context for understanding and judging comparative quality.

- o Potential sources of valid student performance data exist in ongoing state assessment programs, for instance, but

investigations of means for aggregating such information are only just underway for state by state comparisons. The importance of test content receives less attention.

- The idea of a national test to estimate overall national system performance recurs periodically, with the National Assessment of Educational Progress the current version of the idea. Scant attention has been paid to costs and benefits of linking existing assessment systems to create national indicators.

7. Concern for student achievement and the quality of American education escalates each time an international comparison of student performance is conducted. Yet there has been little consideration of the use of international studies, or the measures generated by them, as benchmarks to protect America's ability to compete in technological, academic, and economic futures. For example:

- The Second International Mathematic Study provided a comparison of the United States and 20 other countries. Results show that the United States performed relatively poorly in comparison with Japan. Less serious consideration was given to the meaning of these data with respect to the role that content coverage, the quality of instruction, or the differences in background, abilities, and attitudes might play in the highlighted performance differences, although data are available on these student and instructional characteristics are available.

8. Because different types of decisions (e.g., policy, institutional, instructional, counseling) require different types of information, a patchwork system for collecting information has been created. Not only are the testing and evaluation procedures used unnecessarily intrusive, but the information produced is overly redundant. The redundancy may be particularly acute for special populations. For example:

- Children participating in a Chapter I program at a midwestern school must take the CTBS in the fall and again in the spring, in addition to mandated state assessment tests, a districtwide norm-referenced test, and an array of curriculum-embedded tests. The information from these tests is never integrated is largely redundant, and only tangentially influences teaching practices.

Problems related to utility and impact.

9. Student testing programs on which much of evaluation depends, are externally imposed, from the top-down, but the use of data for local school improvement is a bottom-up proposition, local and specific in nature. The result is data of limited utility for teachers and school administrators. For example:

- Extensive interviews with district administrators, principals, and teachers in one midwestern school district found that while each of these groups believed the tests had value for the system as a

whole, each group also said the tests were not germane to its own needs. Thus, district administrators said that tests were helpful to teachers; teachers thought them useful to principals and principals felt they were essential to district administrators. In short, no group acknowledged that it found such information valuable.

- o According to a national study of teachers' use of testing, teachers reported very little practical decisionmaking based on formal testing because of the mismatch of test content and instruction, poor reporting formats, and inappropriate timing of results.

10. Schools are supposed to be vehicles of social mobility and equity, giving all students an opportunity to achieve and to reap the benefits of productive participation in society. Although rigorous testing systems are supposed to contribute to this process, evidence suggests that testing may actually impede social mobility. For example:

- o According to a prestigious national study of schooling, testing has contributed to the tracking of students into rigid vocational and academic lines, thereby reducing the prospects for individual growth and satisfaction.

- o The treatment of special populations (e.g., children from different language backgrounds or with different developmental histories) often amounts to placement in dead-end tracks with little opportunity for change or advancement.

11. Tests and evaluation are regarded not only as processes for assessing educational quality, but as significant interventions in themselves that will promote excellence and high standards. There is widespread belief that the imposition of testing systems will focus and motivate learning, but other effects contrary to excellence may also accrue. For example:

- o One eastern school district, echoing teachers' concerns in a national study, reported substantial narrowing of the curriculum, away from science, art, history and higher level skills and toward the basic skill areas assessed on mandated tests.

- o Districts around the country are investing resources to train children in test-taking that could be allocated to encouraging subject matter learning; teaching to the test is a common occurrence.

- o Acceptable pass rates are a political necessity, resulting in cut-scores that reflect neither excellence nor even minimum competency.

These three problem clusters, quality of information, quality of inferences and interpretation, and utility and impact of testing and evaluation reforms are central to our problem-focused R&D program. Although better instruments, better interpretations and better understandings of the consequences of testing and evaluation are demanded, the need runs much deeper. The problems are social and epistemological as well as technical.

Problem-focused Research Programs

The conceptual framework defining the CSTES research agenda reflects these perspectives, emphasizing the role of information in improving educational quality and the need for better information about educational quality to facilitate that improvement process. The three research programs derived from the framework reflect areas where significant problems exist in practice and where both steady and identifiable progress can be made.

1. The Testing for the Improvement of Learning Program (Testing) focuses research attention on the design of measures of student learning processes and achievement so that test information can be used to improve instruction and performance. The program emphasis is on improving the quality and validity of measures of student performance and their utility in meeting students' instructional needs. Conceptual syntheses, theoretically-based empirical studies and exploratory research and development of content based measures at the precollegiate and postsecondary levels are planned. These projects address the primary program objective of improving the validity of student performance measures by: improving the content base of measures; improving the usefulness of measures for multiple instructional purposes; broadening approaches to assessing student performance to increase their fairness and utility; integrating research in human cognitive processing and in assessment; and exploring the applications of technology for test development, administration, and analysis.

2. The Systems for Evaluating and Improving Educational Quality Program (Evaluation) is designed to strengthen methodologies for using evaluation to improve educational quality. It seeks to decentralize evaluation systems to the local school level where they can help teachers and school administrators to understand their problems and better meet the instructional needs of students while at the same time accommodating the information needs of local and state policymakers. Conceptual synthses, field-based empirical studies, and research and development projects are proposed to accomplish the primary program objective: To improve the validity of inferences about educational quality by developing methodologies for articulating information needs at the various levels of the educational system; by expanding the band of indicators used to understand and judge quality; by integrating a variety of measures to provide a better picture of educational quality at the precollegiate and postsecondary levels; by exploring the organizational and technical requirements for multilevel evaluation systems; and by conducting analyses of the conceptual and theoretical underpinnings of the evaluation process.

3. The Impact of Testing and Evaluation on Educational Standards, Policy and Practice Program (Impact) seeks to examine the actual effects of testing and evaluation on educational quality and their role in promoting excellence and equity. The program will also monitor and analyze the implementation and quality of new test and evaluation developments on the national level, particularly as they serve as measures of educational reform. The results of the program will provide significant information for educational policymakers responsible for the design of educational programs. The program also is designed to assess and facilitate the impact of CSTES research and development on educational policy and practice and to serve a needs assessment function for future R&D.

The three programs are designed to interact and to support the underlying reason for a center of research and development rather than support for individual products. Explorations in the Testing Program will influence the types of measures used in the evaluation systems studied in the Evaluation Program. Feedback about effects or identification of promising practices obtained in the Impact Program can affect both goals and research plans in both the Testing and Evaluation Programs. Productive findings in the Testing and Evaluation Programs should, in the long run, show their effects in the work of the Impact Program.

Planned institutional function activities incorporate a number of strategies to assure that the results of the research programs are widely disseminated to intended audiences -- teachers; school, district, and state administrators; state and local policymakers; test publishers; and other researchers -- and that they influence future educational research, policy and practice.

References

Alkin, M. Evaluation theory development. Evaluation Comment, 1969, 2(1), 1-4.

Alkin, M., Daillak, R., & White, P. Using evaluations: Does evaluation make a difference? Beverly Hills, CA: Sage Publications, 1979.

Alkin, M., Jacobson, P., Burry, J., Ruskus, J., White, P., & Kent, L. A guide for evaluation decision makers. Beverly Hills, CA: Sage Publications, 1985.

Baker, E. A framework for integrating testing and instruction in school districts. Paper presented at the annual meeting of the Evaluation Research Society, Austin, TX, October, 1981.

Baker, E.L. Evaluating educational quality: A rational design. Invited address, Educational Policy and Management, University of Oregon, 1983.

Baker, E.L. Educational reform in the 80's: Prospects and problems. Invited presentation to the National Conference of the National Educational Association's Blue Ribbon Task Force on Educational Excellence, Washington, D.C., October 1984.

Bank, A., & Williams, R.C. The search for consequences: Assessing the impact of district instructional information systems. Evaluation and Program Planning, Fall 1984a, 6(3).

Bank, A., & Williams, R.C. School district use of testing and evaluation for instructional decision making. CSE Report No. 204. Los Angeles, CA: UCLA Center for the Study of Evaluation, 1984b.

Berman P., & McLaughlin, M. W. Federal programs supporting educational change Vol III: Implementing and sustaining innovations. R-1589/8-HEW, Santa Monica, CA: Rand Corporation, 1977a.

Berman, P., & McLaughlin, M.W. Federal programs supporting educational change, Vol. 7: Factors affecting implementation and continuation. Prepared for the U.S. Office of Education, Dept. of Health, Education, and Welfare, 1977b.

Bock, R.D. Contributions of empirical Bayes and marginal maximum likelihood methods to the measurement of individual differences (September 1984). To appear in the Proceedings of the 23rd International Congress of Psychology, Acapulco, Mexico, 1985.

Bock, R.D., & Aitkin, M. Marginal maximum likelihood estimatation of item parameters: Application of an EM algorithm. Psychometrika, 1981, 46, 443-459, 47,345 (errata).

Bock, R.D., Mislevy, R.J., & Woodson, C. The next stage in educational assessment. Educational Researcher, 1982, 11,4-11, 16.

Boyer, E.L. High School: A Report on Secondary Education in America, New York: Harper & Row, 1983.

Brown, J.S., & Burton, R.R. Diagnostic models for procedural bugs in mathematics. Cognitive Science, June 1984, 2, 155-192.

Bryk, A. (Ed.). Stakeholder-based evaluation. New Directions for Program Evaluation. Vol. 17. San Francisco: Jossey-Bass, 1983.

Burstein, L. Analysis of multilevel data in educational research and evaluation. In D. Berliner (Ed.), Review of Research in Education, Vol. 8. Washington, DC: American Educational Research Association, 1980, 158-233.

Burstein, L. Investigating social programs when individuals belong to a variety of groups over time. CSE Report No. 173. Los Angeles, California: UCLA Center for the Study of Evaluation, 1981.

Burstein, L. Using multilevel methods for local school improvement: A beginning conceptual synthesis. Report to NIE (NIE-G-80-0112, P3). Los Angeles, CA: UCLA Center for the Study of Evaluation, 1983.

Burstein, L., Baker, E.L., Aschbacher, P., & Keesling, J.K. Using state test data for national indicators of educational quality: A feasibility study. Los Angeles, CA: UCLA Center for the Study of Evaluation, 1985.

Burstein, L. & Guiton, G. Methodological Perspectives on Documenting Program Impact. In B. Deogh (Ed.), Advances in Special Education, Vol. 4. Greenwich, CT: JAI Press, Inc., 1984, 21-42.

Burstein, L., Schwille, J., Travers, K., Robitaille, D.F., Cooney, T., & Rolsin, D. Second international mathematics study: Student and classroom process in early secondary school. London: Pergamon Press (in press).

Clark, B.R. The school and the university: An international perspective. Berkeley, CA: University of California Press, 1985.

Cook, T.D., & Campbell, D.T. Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally & Co., 1979

Cooley, W., & Bickel, W. Evaluation use: Pittsburgh case studies. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, April 1985.

Cronbach, L.J. Designing evaluations of educational and social programs. San Francisco: Jossey-Bass, Inc., 1982.

Cronbach, L.J., et al. Toward reform of program evaluations. San Francisco. CA: Jossey-Bass, 1980.

Curtis, M.E., & Glaser, R. Reading theory and the assessment of reading achievement. Journal of Educational Measurement, 1983, 20, 133-147.

Dorr-Bremme, D. Assessing students: Teachers' routine practices and reasoning. Evaluation Comment, 1983, 6(4), 1-12.

Elliott, E.J., & Hall, R. Indicators of performance: Measuring the educators. Educational Measurement: Issues and Practice, 1985, 4(2), 6-8.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-21.

Glass, G. Educational knowledge use. The Educational Forum, 1972.

Goodlad, J.I. A place called school. New York: McGraw Hill, 1983.

Green, B., Bock, R.D., Humphrey, L.G., Lynn, R.L., & Rekase, M.D. Technical guidlines for assessing computerized adaptive testing. Journal of Educational Measurement, 1984, 21, 347-360.

Haertel, E., & Calfee, R. School achievement: Thinking about what to test. Journal of Educational Measurement, 1983, 20(2), 119-132.

Hathaway, W.E. Hopes and possibilities for educational information systems. Presented at the invitational conference "Information Systems and School Improvement: Inventing the Future," UCLA Center for the Study of Evaluation, Los Angeles, February 1985.

Hayes, J.R., & Flower, L.S. Identifying the organization of writing processes. In L.W. Gregg & E.R. Steinberg (Eds.), Cognitive processes in writing. Hillsdale, NJ: Erlbaum, 1980.

Herman, J. Local evaluation and future possibilities. Educational Evaluation and Policy Analysis, 1985.

Herman, J., & Dorr-Bremme, D. Uses of testing in the schools: A national profile. New directions for testing and measurement, No. 19. San Francisco: Jossey-Bass, September 1983 (pp. 7-17).

House, E.R. The logic of evaluative argument. Monograph No. 7. Los Angeles: UCLA Center for the Study of Evaluation, 1977.

House, E. R., Glass, G. V. McLean, L. D. & Walker, D. F. No simple answer: A Critique of the follow through evaluation. Harvard Educational Review, 1978, 48(2), 128-160.

Howe, H. The value of research and disciplined inquiry to the improvement of education. In R. Glaser (Ed.), Improving Education: Perspectives on Educational Research. Pittsburgh, PA: National Academy of Education, 1984.

Larkin, J., McDermott, J., Simon, D., & Simon, H. Expert and novice performance in solving physics problems. Science, 1980, 208, 140-156.

Lindblom C.E. & Cohen D.K. Usable knowledge: Social science and social problem solving. New Haven: Yale University Press, 1979.

Lord, F.M. Applications of item reponse theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

Pace, R. Measuring outcomes of college. San Francisco: Jossey-Bass, 1979.

Pelz D.C. Innovation complexity and the sequence of innovating stages. Knowledge: Creation, Diffusion, Utilization, 1985, 6 (3), 261-291.

Purves, A. The IEA Study in Written Composition. Urbana: University of Illinois, 1980.

Quellmalz, E. Designing writing assessments: Balancing fairness, utility and cost. Educational Evaluation and Policy Analysis, 1984, 6(1).

Reisner, E., Alkin, M., Boruch, R., Linn, R., & Millman, J. Assessment of the Title I evaluation and reporting system. Washington, D.C.: U.S. Department of Education, Office of Planning, Budget and Evaluation, April 1982.

Scriven, M. The Methodology of evaluation. In R.W. Tyler, R.M. Gagne, & M. Swiven (Eds.), Perspectives of curriculum evaluation. AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally, 1967, pp. 39-83.

Scriven, M. Evaluation perspectives and procedures. In W.J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, CA: McCutcheon, 1974.

Shavelson, R.J. Evaluation of nonformal education programs: the applicability and utility of the criterion-sampling approach. Oxford: Pergamon Press, 1985a.

Shavelson, R.J. The measurement of cognitive structure. Paper presented at the annual meeting of the American Educational Research Association, Symposium on the "Psychology of Learning Science", April 1985b.

Shepard, L.A., Camilli, G., and Averill, M. Comparison of procedures for detecting test-item bias with both and external ability criteria. Journal of educational statistics, 1981, 6, 317-375.

Sirotnik, K.A., & Burstein, L. Making sense out of comprehensive school-based information systems. Los Angeles: UCLA Center for the Study of Evaluation, 1984.

Sirotnik, K.A., & Burstein, L. Measurement and statistical issues in multilevel research on schooling. Educational Administration Quarterly, 1985.

Sirotnik, K.A., Burstein, L. & Thomas, C. Systemic evaluation. Report to NIE (NIE-G-83-0001). Los Angeles, CA: UCLA Center for the Study of Evaluation, 1983.

Sizer, T.R. Horace's compromise: The dilemma of the American high school. Boston: Houghton Mifflin, 1984.
York: Macmillan, 1983.

Smith, M.S. A Framework for the Development of National Educational Indicators. Prepared for the consideration of the Council of Chief State School Officers, 1984.

Stake, R.E. The case-study method in social inquiry. Educational Researcher, February 1978, 7, 7-8.

Stufflebeam, Daniel, L, Faly, W.J., Guba, E.G., Hammard, L.R., Merriman, H.O., and Provus, M.M. Educational evaluation - decision making in education. Itasca, IL: Peacock, 1971.

Tatsuoka, K.K., & Tatsuoka, M. M. Detection of aberrant response patterns and their effect on dimensionality. Journal of Educational Statistics, 1982, 7, 215-231.

Travers, K. The international mathematics curriculum intention & implementation. London: Pergamon Press (in press).

Weick, K. Educational organizations as loosely coupled systems. Administrative Science Quarterly, March 1976, 2.

Weiss, C.H. Evaluation research. Englewood Cliffs, NJ: Prentice Hall, 1972.

Weiss C.H. (Ed.). Using social research in public policymaking. Lexington, MA: D.C. Heath, 1977.

Williams, R. C., & Bank, A. The search for consequences: Assessing the impact of district instructional information systems. Educational Evaluation & Policy Analysis, 1984, 6(3), 267-282.

Wirtz, W., & LaPointe, A. Measuring the quality of education: A report on assessing educational progreses. Washington, DC, 1982.