DOCUMENT RESUME

ED 278 712 TM 870 142

AUTHOR Stufflebeam, Daniel L.
TITLE [Center on Student Testing, Evaluation, and Standards Setting.] Research and Development Mission. Final Performance Report.
INSTITUTION Western Michigan Univ., Kalamazoo.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Sep 85
GRANT PA-84-3
NOTE 126p.
PUB TYPE Reports - Descriptive (141) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS *Academic Standards; Adaptive Testing; Computer Assisted Testing; Consortia; *Educational Testing; *Long Range Planning; Personnel Selection; Program Development; *Program Proposals; Psychometrics; *Research and Development Centers; Research Needs; Test Results; Test Use
IDENTIFIERS Boston College MA; *Center on Student Testing Evaluation and Standards; Dallas Independent School District TX; University of Kansas; Western Michigan University

ABSTRACT

A consortium of four institutions, based at Western Michigan University, was to plan a Research and Development Center on Student Testing, Evaluation, and Standards Setting. The four consortium sites are Boston College, the Dallas Independent School District, the University of Kansas, and Western Michigan University. This report is a final performance report for the planning grant made to Western Michigan University by the National Institute of Education. The first section provides a summary of the planning activities actually conducted under the grant. It includes a description of particular problems and successes, a list of participants and their affiliations, and an evaluation report for the planning project. The second section is a technical report on the Research and Development (R and D) mission for the Center. It updates the mission and strategy statement contained in the planning grant application and includes an agenda, with supporting justification for R and D within the Center's mission. The third section contains a futures paper, as required by the grant announcement. It is intended for practitioners and researchers and describes work needed in the mission area of the Center to accomplish desirable goals by 1990. A 16-page bibliography of sources used in the planning project and supporting documentation for particular parts of the planning project are appended to this performance report. (JAZ)

FINAL PERFORMANCE REORT ON RESEARCH AND DEVELOPMENT MISSION

A final report submitted to the National Institute of Education in compliance with Grant Announcement No. NA-84-3

Submitted by
Western Michigan University
on behalf of itself and

Boston College
Dallas Independent School District
University of Kansas

Daniel L. Stufflebeam
Principal Investigator

Evaluation Center
Western Michigan University
Kalamazoo, Michigan 49008
(616) 383-8166

FINAL PERFORMANCE REPORT

A final report submitted to the National Institute of Education in compliance with Grant Announcement No. PA-84-3

Submitted by
Western Michigan University
on behalf of itself and

Boston College
Dallas Independent School District
University of Kansas

Evaluation Center
Western Michigan University
Kalamazoo, Michigan 49008
(616) 383-8166

September 1985

CONTRIBUTORS

Compiled by:

James Sanders, WMU
Gregory Hagerman, WMU

Based on the written contributions of:

Gilbert Austin, U. Maryland
William Cooley, U. Pittsburgh
Ronald Crowell, WMU
Richard Frisbie, WMU
Douglas Glasnapp, U. Kansas
Egon Guba, Indiana University
Ronald Hambleton, U. Massachusetts, Amherst
Walter Haney, Boston College
Theresa Hollowell, WMU
Yvonna Lincoln, U. Kansas
George Madaus, Boston College
Robert Mendro, Dallas Independent School District
Joseph Pedulla, Boston College
Daniel Stufflebeam, WMU
Lyke Thompson, WMU

TABLE OF CONTENTS

I. Summary of Planning Activities and Evaluation Report

A consortium of four institutions, based at Western Michigan University, was established to plan a Research & Development Center on Student Testing, Evaluation, and Standards Setting. This consortium is committed to a long period of collaboration on research and development addressing important issues in testing, evaluation, and standard setting. The four consortium sites are Boston College, the Dallas Independent School District, the University of Kansas, and Western Michigan University.

This report is a final performance report for a planning grant made to Western Michigan University by the National Institute of Education. The report is required to fulfill conditions of the grant, as described in Grant Announcement No. PA-84-3. This first section provides a summary of the planning activities actually conducted under the award to Western Michigan University. It includes a description of particular problems and successes, a list of participants and their affiliations, and an evaluation report for the planning project. The second section is a technical report on the R and D mission for the Center. It updates the mission and strategy statement contained in the planning grant application and includes an agenda, with supporting justification for R & D within the Center's mission. The third section contains a futures paper, as required by the grant announcement. It is intended for practitioners and researchers and describes work needed in the mission area of the Center to accomplish desirable goals by 1990. A bibliography of sources used in the planning project, and supporting documentation for particular parts of the planning project are appended to this performance report.

The consortium has an exemplary record in working with state education agencies (SEAs) and local education agencies (LEAs) in furthering educational development derived from integrated research. Western Michigan University's (WMU) Evaluation Center, under the direction of Drs. Stufflebeam, Sanders, and Bunda, has a long and successful history of R & D at SEAs and LEAs and, in particular, the development of minority and special client programs. Boston College's (B.C.) Center for the Study of Testing, Evaluation, and Educational Policy (CSTEEP), under the leadership of Dr. Madaus, became seriously involved with policy issues through an extensive process of research at local, national, and international levels. Dallas Independent School District's (DISD) Department of Research, Evaluation and Information Systems, under the leadership of Dr. Webster, is home for the premier local educational agency R & D center in the world. The University of Kansas' (U of K) Center for Educational Testing and Evaluation, directed by Drs. Poggio and Glasnapp, has performed a state government mandated assessment and development of student testing throughout this heartland state. The consortium has also secured the involvement of two top psychometricians (Drs. Hambleton and Swaminathan from the University of Massachusetts at Amherst) to add to the scope of its R & D efforts.

The objectives of the planning grant were to:

1. **Create the governance structure, the administrative structure, and the formal agreements to undergird the consortium**

2. **Explicate the proposed mission and immediate and long term objectives**

3. **Develop and reach agreement on an agenda of projects**

4. **Develop staffing facilities and work plans for the initial effort**

5. **Act affirmatively in the recruitment of minorities and women to key staff positions in the Center**

6. **Develop a network of cooperating agencies to help plan, conduct, and use the Center's projects**

7. **Develop a plan and arrangements for an ongoing evaluation of the Center**

The first goal of the consortium was to secure letters of institutional support and formal agreement from each member institution. This was done concurrently with the development of the formal governance structure since the same people, university presidents, had to be contacted. Having achieved this commitment, the director, Dr. Stufflebeam, convened the consortium members and school district superintendent in Chicago following the annual American Educational Research Association/National Council on Measurement in Education (AERA/NCME) convention to discuss the specifics of administrative structure. It was there that the decision to incorporate a deputy director, Dr. Thompson, into the administrative structure was made to strengthen the Center's day-to-day administration. The identification of an executive committee, the individual site directors (Drs. Madaus, Poggio, Hambleton, Webster, and Sanders), to share in the administrative responsibilities of the Center helped to create an atmosphere of collegiality and trust.

The successful operating of the Center was seen to be contingent upon the establishment of a capable system of governance and administration. Administration would be responsible for the operations and quality of R & D products. It would coordinate these activities and provide for the temporal and fiscal efficacy of effort and product. It would insure that the needs of all concerned populations were addressed. Central to the development of administrative expertise is the requirement of an ongoing evaluation of R & D and all of its attendant activities. The evaluation component of this consortium will provide information to its national advisory panel to assist it in directing and redirecting the thrust and scope of its R & D activities. The evaluations will, also, be instrumental in determining the success of the consortium in meeting its interim and ultimate objectives in disseminating results. External evaluations will serve to validate the internal evaluation process and further the practice of relevant R & D. Together, the National Advisory Panel, the Governance Board, and both internal and external evaluators will work to insure that an exemplary standard of ethical practice in research, development, and dissemination is adhered to. The Governance Board will participate in the inevitable political decisions that must occur in any cooperative enterprise. Its function will involve the authoritative allocation of resources across programs to insure a viable collaborative organization. Consortium site representation will be equal and institutional responsibility and support enhanced by the Board. The governance and administration structure created to achieve the first objective of the planning grant is provided in Figures 1 and 2.

Figure 1. General Structure of the Center

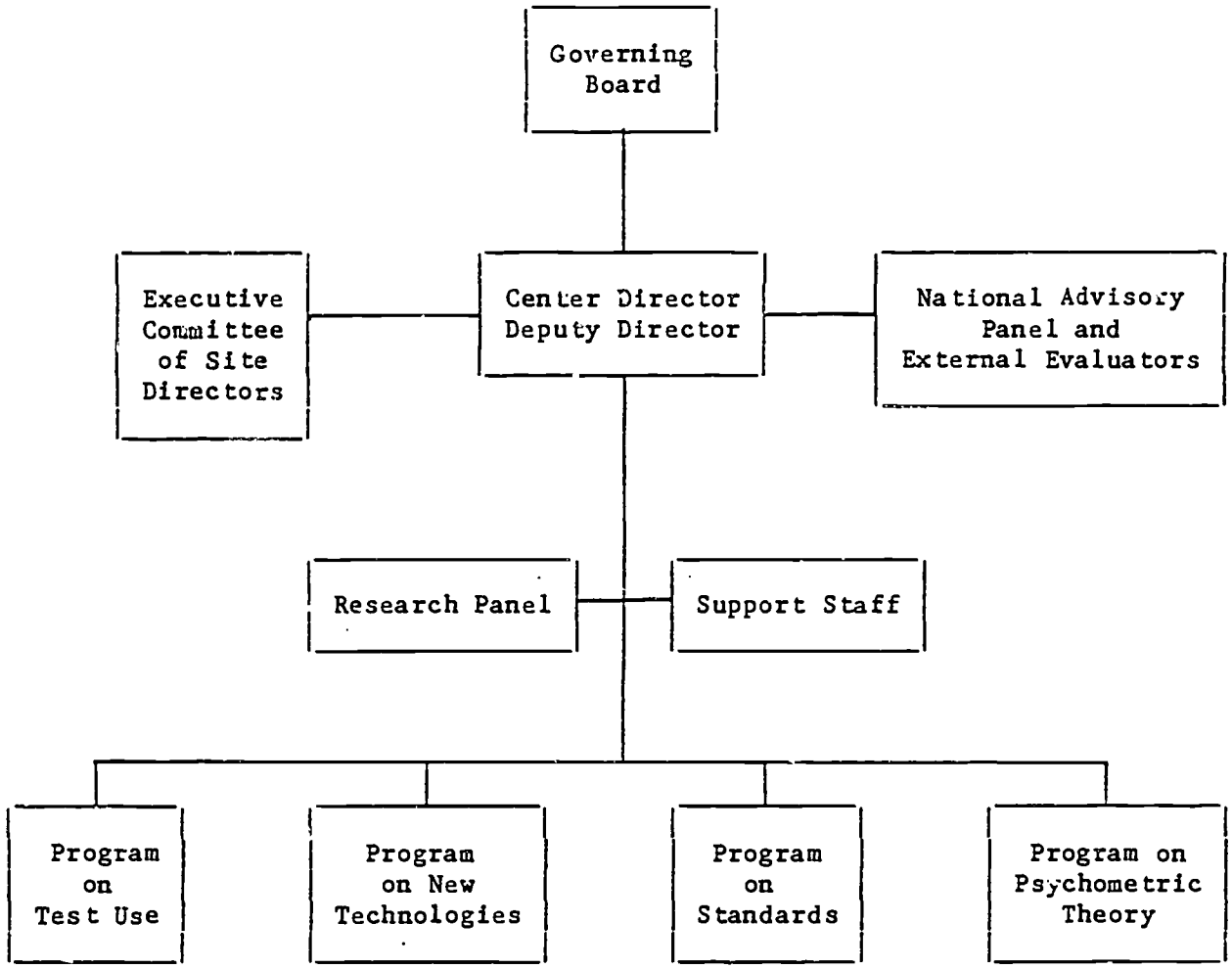
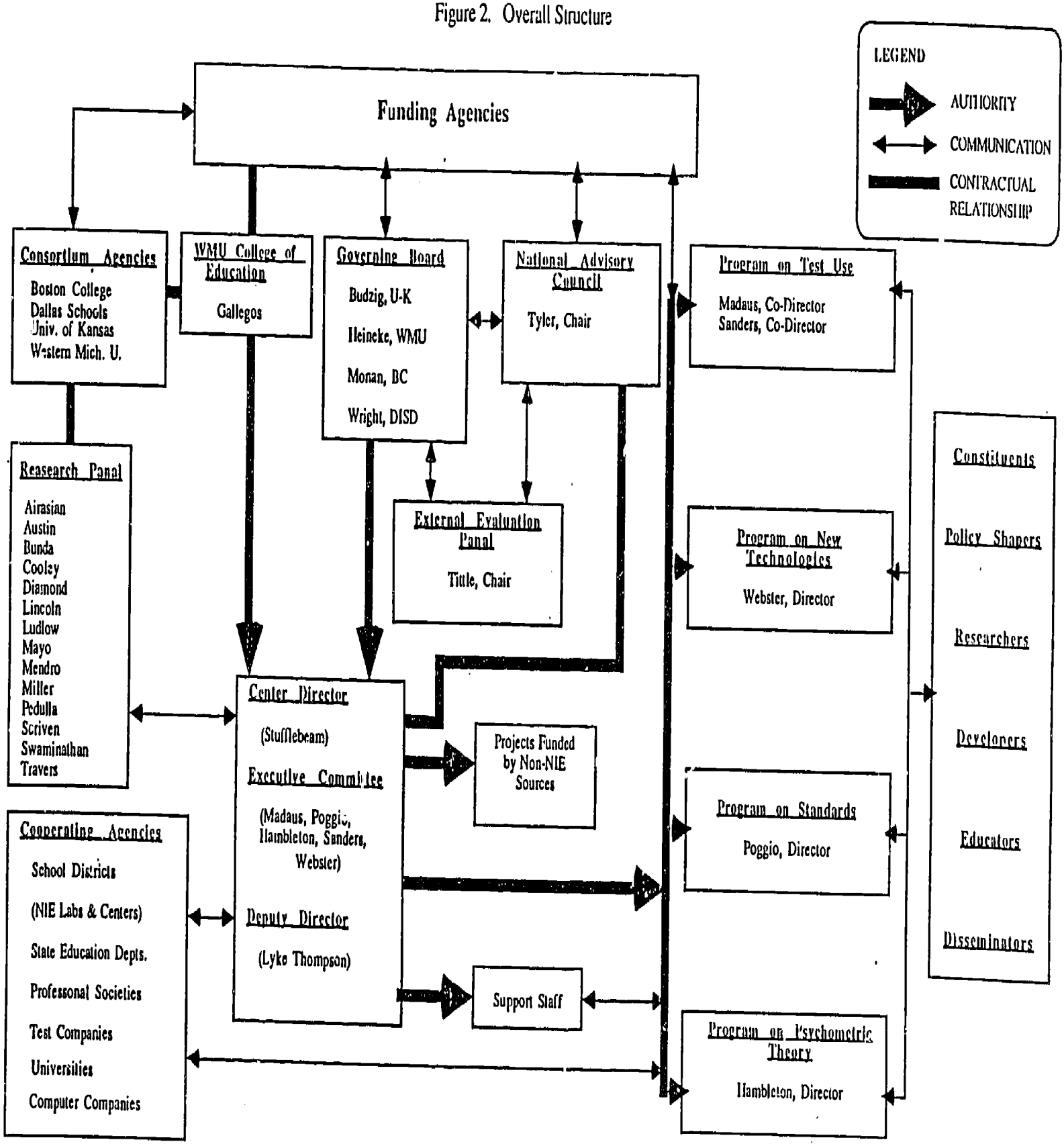


Figure 2. Overall Structure



The planning committee, composed of the director, deputy director, executive committee, and research panel was then formed to produce and justify the Center mission and its short and long term goals. The research panel consisted of researchers from the member institutions, cooperating agencies, and independent agencies. A needs assessment, in the form of a key informant survey, was conducted to insure the relevance and efficacy of the proposed Center mission. Respondents included officials from such organizations as the National Urban League, Council for Basic Education, American Association of School Administrators, Mexican American Legal Defense Fund, Southwest Educational Development Laboratory, Public Education Association, Center for Teaching & Learning, National Education Association, as well as a number of SEA and LEA practitioners. An analysis of responses from this survey was paired with a content analysis of the original proposal to best find a match between what was needed and what the consortium could best accomplish (See Appendices 1 through 6). The result was an extensive array of problem sets. From this array was chosen a group of programs that addressed the needs that emerged from planning activities.

The consortium examined needs of the educational community that pertained to the Center's proposed mission. This was done through a review and analysis of literature, and through interviews and a survey of key informants. The interviews were exceptionally influential in the conceptualization of needs for R & D on student testing, evaluation, and standards setting. The literature review was performed by staff personnel best qualified in the areas to be addressed.

The next step in this planning process was the submission of position papers dealing with the most critical aspects of the programs. These concept papers were offered by consortium staff members and research panel members to the planning committee as what each researcher viewed as the most important problems and appropriate research and development projects. The papers were distributed to each site and discussed via CompuServe and twice monthly conference calls. Approximately twenty different concept papers were developed in this manner for consideration by the planning committee. The papers went through a number of drafts and revisions by the original authors, as well as by members of the planning committee. These papers were the basis for creating a program agenda. A summary of the concept papers is appended to this report.

A second meeting of key personnel was held at the Evaluation Center at WMU in July to make critical decisions for the final proposal. At this meeting a great deal of work was done in finalizing the agenda of proposed projects and allocation of resources. The ability of the consortium to perform effectively and efficiently was demonstrated at this time. If ever there was a time for a breakdown of process to occur it was during this crucial stage of negotiation in setting priorities. The effectiveness of the organizational structure of this collaboration was apparent in the deliberative process and in the resulting agenda of projects that was established. Agreement on the agenda was brought about by orderly processes set in place through the establishment of the executive committee.

There were three distinct selection processes which took place during the term of this planning grant; 1) **program selection** and 2) **project selection**, described above, and 3) **personnel selection**, to follow.

The selection of R & D personnel was an organic process of melding a diverse group of experts with unique concerns and perspective into a synergistic community of educational scholars with a broad purview and common goals. To achieve this complex organizational structure, an alliance of all members had to be established. This alliance had to provide for the individual staffing needs of site projects and at the same time offer the Center director the ability to make budget and staffing projections and to maintain the cohesiveness of the Center's R & D programs.

It was initially the responsibility of Dr. Stufflebeam, the proposed Center director, to compose a team of researchers able to achieve the goals set forth in the RFP. To this end he invited colleagues who had a history of productive R & D experience, and previous collaboration experiences to respond to WMU's proposal for a planning grant for the R & D center. Of those responding it was determined that the greatest strength of past accomplishment and present capability was to be found at Boston College, the Dallas Independent School District and the University of Kansas. Since the inception of the proposal process, Dr. Madaus at B.C., Drs. Denton and Webster at DISD, Dr. Poggio at U of K, and Dr. Stufflebeam have labored to bring together that team of researchers best qualified to achieve the R & D Center goals. Very early, it became obvious psychometric theory input would be of fundamental importance in the Center's work, and Drs. Hambleton and Swaminathan at the University of Massachusetts, Amherst, were added to the team roster. Each member of the total team has a history of excellence and delivery of research products. Their institutions provided letters and pledges of institutional and financial support. This includes the commitment of additional key personnel from each member institution.

The determination of this group to create and sustain an excellent research organization led them to recruit, not only, the brightest lights at their own organizations but the best available talent throughout the country to augment and enhance the consortium's effort. They asked Dr. Tyler to serve as head of the National Advisory Panel because it was believed that Dr. Tyler's unmatched experience in educational R & D related to student testing would be invaluable in providing advice to the consortium in area's both practical and esoteric. Dr. Tittle was also sought to contribute her particular expertise in evaluation and concerns of special populations to the chair of the external evaluation panel. The significance of these two appointments is reflected in the national prominence these two hold in the field of educational R & D related to testing. Their appointments also bolstered the ability of the consortium to plan significant and technically sound programs of research.

A great many other notable individuals have concurred with the objectives set forth by the consortium and have pledged to assist in planning and implementing projects. These include Drs. Guba, Austin, Cooley, Diamond, Iverson, Mayo, Scriven, Sandberg, Pullin, Merwin, Wallace, Lincoln, Travers, and Rao, who have each contributed substantial time and resources to the project thus far. Letters of cooperation, communication, and collaborative intent were solicited and received from sixteen recipients of NIE R & D Center planning grants, six recipients of NIE Laboratory planning grants, fourteen national and international consultants and R & D organizations, and about fifty school districts.

While the individual plan of action in the planning stage of this proposal was developed on site, the finished product (the proposal) reflects a united effort by all the consortium personnel and a considerable amount of support from outside but interested parties. This support at WMU included, but was not limited to, Mr. Wil Emert, Division of Research and Sponsored Programs; Mr. Frank Jamison, Media Services; Mr. James Kirklin, Academic Services Technical Support; Ms. Jane Bauer, Office of Conferences and Institutes; and Ms. Janice Argue, Grants and Contracts.

Affirmative action plans and records are on file at the WMU Evaluation Center for Boston College, the Dallas Independent School District, the University of Kansas, the University of Massachusetts, and Western Michigan University. Each institution is formally committed to policies of recruitment and appointment that provide equal employment opportunity to all qualified persons, in all job classifications, in recruiting, hiring, employment upgrading, promotion, selection for training, transfer, termination, compensation, benefits, and working conditions, without regard to race, color, religion, age, sex, national origin, veteran status, or handicap. The consortium members have developed and identified strategies necessary to recruit key staff people not named in the proposal. These staff positions will be filled in accordance with the letter and spirit of all civil rights and equal employment opportunity legislation, paying particular attention to the hiring and professional development of minorities, women, the handicapped, and the elderly. The positions to be staffed include key line staff posts, members of the National Advisory Panel, consultants, visiting scholars, collaborators in research, and support staff.

The construction of governance and administration policy was of major import in the development of a strong R & D organization. While the consortium recognized the need for this capability it still felt that its investigatory efforts should be focused on R & D. It quickly became apparent that the administrative component was necessary and complementary to the R & D mission. What emerged was a strategy for allocating resources primarily for R & D, and secondarily for activities involving governance, evaluation, and administration. These activities became support mechanisms to the R & D, collaboration and dissemination efforts.

Fiscal and physical resources were assessed at each site in light of their ability to contribute to the collaboration/dissemination effort. Papers on previous collaborative R & D organizations were studied and discussed. Professionals in the field were consulted for their insights and inputs. Practicing educators were requested to add their ideas about R & D strategy. From this information emerged a strategy of dissemination involving comprehensive collaboration, communication, and professional development relating to educational research and development. This collaboration includes a network of cooperating agencies already taking form as noted above. Those agencies or individuals who either have sent letters of collaborative intent or who participated in the key informant survey have certainly influenced the scope and direction of research and development proposed by this consortium. In this manner it may be said that they have already assisted in the planning of the Center's programs and projects and they have in their correspondence and in personal communications indicated their willingness and desire to participate in the conduct of research and its subsequent application and evaluation.

The seven objectives of the planning grant were evaluated according to the following criteria:

1. **Management of the project**--was it carried out as designed, in a timely and cost-efficient manner, with reports and deliverables submitted as planned?

2. **Scope of the project**--were the appropriate individuals and groups involved in a meaningful way in the planning of the Center? Were the sources of information appropriate for the information that was needed? Were any shortcomings in input, due to procedures or sources that were employed, rectified?

3. **Use of information in planning the Center**--was information gathered during the planning process used in developing the final plan? Can the final plan be justified using the available planning information?

4. **Fiscal responsibility**--was the grant support used to support important planning work that otherwise might not have been feasible?

The evaluation process for the planning grant involved extensive needs assessment (context evaluation), and reviews of drafts designs, concept papers, and draft sections of the proposal for an institutional grant (input evaluation). The context evaluation included an extensive review of pertinent literature which was reported in the proposal for an institutional grant and summarized in the second section of this final performance report. A bibliography of all sources found to have some bearing on the needs for this Center is also appended to this report. The context evaluation also included interviews with educators and researchers about needs that should by addressed by the national center. The interviews were reported in the form of memoranda to the files which were then used by the planning team to conceptualize proposed programs and projects. Finally, the context evaluation included a survey of key informants to obtain their perceptions of need and priorities for the Center. The results of this survey are provided in Appendix 3 of this final report. The key informants were selected because of their knowledge of issues faced by traditionally under represented groups and of issues that reflected national concerns. The context evaluation resulted in the identification of needs that were reviewed by the planning team. These needs became the basis for program planning.

Input evaluation was done at each step of the planning process. Alternative designs were solicited from key personnel and consultants for each component of the later plan. The planning team then served as evaluators who submitted their recommendations on designs, program specifications, and drafts of the institutional grant proposal to the Center director. In addition, content analyses of reviews of the WMU planning grant applications and of Grants Announcement No. PA-84-3 were performed to guide design decisions. Appendices 5 and 6 of this report reflect the findings of the content analyses.

The director, Dr. Stufflebeam, and his staff at WMU then used these evaluations to select or revise each part of the institutional grant proposal. Designs often went through several cycles of this evaluation process. Thus, the institutional grant proposal has been evaluated systematically by key

personnel and consultants. Individuals who were involved in the planning process are listed in Appendix 1. The planning project was carried out as planned with rigorous attention given to identification of needs and to obtaining and getting reviews of a wide range of alternative designs for each component of the institutional grant proposal. Drs. Sanders and Stufflebeam worked closely to assure that the final proposal for the Center:

- was based on the planning process that was described in the WMU application for a planning grant
- used input from the important constituencies listed in the application for a planning grant
- was a valid representation of high priority needs in student testing, evaluation, and standard setting
- made efficient and effective use of planning grant support

The evaluation plan for the proposed Center was also reviewed using the Standards for Evaluations of Educational Programs, Projects and Materials, published by the Joint Committee on Standards for Educational Evaluation.

The operational strategy for the planning project has closely paralleled the strategy that we foresee for the operation of the Center itself. Specifically, we have:

1. Identified critical needs in schools that should be addressed by the Center
2. Involved qualified practitioners, statespersons, and researchers in the planning of long-term programs of research and development to be undertaken by the Center
3. Arranged for school-based research and development settings that represent the diversity of settings to be found in the United States
4. Arranged for collaboration with staff members of local school districts, regional laboratories and other research and development Centers, professional associations, state education agencies, and other relevant educational agencies to accomplish the work of the Center
5. Established communication channels to disseminate the work of the Center

II. Report on Research and Development Mission

In compliance with the requirements of Grant Announcement No. PA-84-3, this technical report on the research and development mission for a proposed NIE Center on Student Testing, Evaluation, and Standard-Setting has been prepared and is hereby submitted as part of the final report for the Western Michigan University Planning Grant. This technical report updates the mission and strategy statement contained in the planning grant application. It also includes an agenda, with supporting justification, for research and development within the Center's mission. This report contains information that was in the Western Michigan University institutional grant proposal. The intended audience for this report is the National Institute of Education with the understanding that NIE may make it available to the recipient of the Center institutional grant and to other researchers and practitioners in the mission area of the Center.

The planning period between January and August, 1985 provided an opportunity to compile research findings within the Center's mission, to collect information about practitioners' needs, and to reflect on the most appropriate allocation of limited research and development resources that would address and build upon our findings.

In this first section of the Technical Report on Research and Development Mission, we discuss the rational for the mission statement provided in our institutional grant proposal. In the second section, we discuss reasons for the selection of the Research and Development approaches that were described in our institutional grant proposal.

Mission for the Center

We recognized from the outset of the planning project that the funds for the proposed center would be limited and that we had to converge on certain needs if the gains from research and development are to have practical significance. Our reviews of past research and our discussions with practitioners led us to focus our plans on the conduct of research and development on testing.

Testing of students has been a part of education for several millennia. Prior to the nineteenth century, testing was usually conducted via oral exam or some sort of performance examination. In the mid-nineteenth century, written exams came to be widely used in the schools of the United States. Then in the late nineteenth century and early twentieth century, with the founding of psychology and the huge growth in the enterprise of schooling, various forms of achievement, psychological, and physiological testing were introduced in the schools of America.

Historically, the main purpose of student testing has been the evaluation of student learning. In recent decades, however, testing in the schools has taken on additional functions. In the 1930s testing was advocated and increasingly used as a source of information for student guidance and selection into college. In the 1950s new testing programs were introduced as a means of identifying exceptional children with special school needs. In the 1960s and 1970s testing came to be widely used as a means of program evaluation. In the 1970s, many new state-wide testing programs were also introduced as a means of ensuring the accountability of schools, to help guarantee that students were

at least minimally competent in the basic skills, and to monitor grade-to-grade promotion and high school graduation. Now in the 1980s new testing programs are widely advocated as instruments of educational reform. Examples are interactive computer tests, content-relevant tests, and curriculum-embedded tests, which are heavily dependent on advances in psychometric testing theory.

In short, student testing in the schools of the United States now serves an extremely wide variety of purposes and functions. Among these are:

- evaluating student learning
- selecting students for placement into special programs or institutions
- evaluating educational programs
- identifying students with special needs
- improving educational standards
- helping to guarantee the minimal competency of students
- guiding instruction
- informing student guidance and counseling
- monitoring grade-to-grade promotion and high school graduation
- helping to ensure the accountability of the schools
- serving as a tool for educational and psychological research
- communicating with parents regarding student progress

There are, of course, many other roles and ways of describing the roles that student testing currently serves in the schools of the United States, but even this brief listing indicates three general points:

1. Student testing is serving many different purposes and functions;

2. The functions and purposes served by testing seem to be increasing over time;

3. Testing is affecting education at many different levels (i.e., through teachers, administrators, students, parents, schools, and school districts via state and national educational policy; the results of testing and the growth of its use are influencing the very ideas and constructs we use to think about schools and education).

Given such diversity, what then should be the mission of the NIE Center on Student Testing, Evaluation, and Standards? Our answer to this question is premised on six key assumptions, derived from our reading of the history of testing and research on testing and from our extensive experience in working with schools to use testing and evaluation to improve teaching and learning. These assumptions are:

(1) Educational research on testing is most likely to be productive in improving schooling if it is conducted in **close collaboration with local schools, teachers, and administrators.** A tremendous amount of experience over the last three decades has shown that one of the biggest problems with all educational research is that of translating its findings into educational practice. One clear remedy to this problem is to involve schools and educational practitioners in research from the start: from its inception to its execution to its interpretation. Thus, one key assumption underlying our proposal is

that research on testing should be conducted collaboratively with schools and school people across the nation. We think the consortium we have put together will facilitate that collaborative effort.

(2) Most of the testing over the last 80 years has been based on what is widely called the classical test model and employs paper-and-pencil multiple-choice testing. Recent technical developments in both test theory and in electronic and video information processing, however, offer much potential to **use new technology to improve many of the functions served by testing.** This will allow us to break free from both the classical test model and the paper-and-pencil multiple-choice format.

(3) Current models of testing using the paper-and-pencil multiple-choice format are so widely used that it surely will be a fairly long time (i.e., at least five years) before new testing technology can significantly affect the testing practices of the majority of schools in the U.S., but the need for improving education is too urgent to rely exclusively on the long-range strategy of using new technologies for improving testing. Another key assumption, therefore, is that the work of the Center must also encompass **research aimed at making currently available test information more useful.**

(4) In planning a research program for an enterprise as large and diverse as that of educational testing, some means of differentiating functions of testing must be employed. A laundry list of different functions served by student testing is not a very economical way of drawing such a distinction. We think that a useful and fairly **fundamental distinction is between external, or policy-oriented testing, versus internal, or school-oriented testing programs.** The external versus internal distinction refers to the locus of initiation for the testing. External testing programs are those, for example, mandated by national or state educational agencies. Such testing programs may also be referred to as policy-oriented because they are providing information for educational accountability programs and for policy makers that transcends individual schools. In contrast, internal testing programs are those undertaken as a matter of discretion by individual school systems. Such internal testing programs may use externally produced tests but also include teacher-made tests, and systematic observation.

There are two main reasons for drawing this distinction. First is the way in which testing tends to be carried out. External or policy-oriented testing programs are carried out on a relatively large scale, and use the paper-and-pencil multiple-choice format and are machine scored. Though this format of testing may also be used in internal testing programs, the latter type of testing also often includes other methods of assessment, such as fill-in-the-blank, short answer, and essay tests. Second is the way in which these two types of testing programs affect teaching and learning. Internal testing programs may affect teaching and learning quite directly, for instance how an individual teacher instructs an individual student. In contrast, external testing programs affect teaching and learning indirectly, for example via changes in formal curricula and school organization.

(5) Testing may be the most visible instrument and indicator of educational standards and the most widely recognized means of student evaluation, but **educational standards are also embodied in many less obvious and less tangible forms.** Examples of these other less tangible embodiments of educational standards are curriculum requirements, teachers' and parents' expectations and judgments regarding student learning, and even the way time is allocated in schools. Given that the ultimate aim of the work of the Center is improved student learning and higher educational standards through improved testing practices, **testing cannot be viewed in isolation from other forms in which educational standards are embodied.**

(6) Given the fact that testing serves so many different functions affecting education on many different levels, and is only one embodiment of educational standards, the **research of the Center must employ multiple methods of research and multiple research perspectives.**

The mission for the Center that was presented in our proposal for an institutional grant flowed from our six key assumptions. The mission statement was as follows:

This Center will contribute to the improvement of the schools of the United States by conducting a collaborative program of research and development directed at improving student testing, evaluation, and standard setting. The Center will directly benefit policy makers, educators, parents, and students by (a) enhancing appropriate and fair use of currently available tests and (b) developing new and improved evaluative procedures, instruments, and systems. The efforts will be focused on needs and problems at both state and local levels.

In light of the key assumptions that we made, the overall mission of the Center will be to conduct research and development on testing:

- in close collaboration with SEAs and LEAs and intensive involvements with selected schools
- by taking advantage of and promoting the potential of new developments in psychometrics and new technologies
- while at the same time seeking to improve the utility of currently available test information
- with regard to both external and internal testing programs
- by viewing testing in conjunction with other embodiments of educational standards
- using multiple methods and perspectives of research

all, with the ultimate goal in mind of improving student learning and raising educational standards.

Strategy for the Center

In our planning grant proposal, we took the position that the Center must involve collaborators with different areas of expertise and experiential background, and must be based on selected areas of need that will be consistently updated. We projected a strategy for the Center that would:

1. Identify critical needs in schools that should be addressed by the Center

2. Involve qualified practitioners, statespersons, and researchers in the planning of long-term programs of research and development to be undertaken by the Center
3. Arrange for school-based research and development settings that represent the diversity of settings found in the United States
4. Arrange for collaboration with staff members of local school districts, regional laboratories and other research and development Centers, professional associations, state education agencies, and other relevant educational agencies to accomplish the work of the Center
5. Establish communication channels to disseminate the work of the Center.

Recognizing that a series of independent research studies within the mission of the Center would be unlikely to yield findings sufficiently focused to effect improvements in schools, we planned to emphasize programmatic research using personnel from across collaborating sites. The programs of research would be consistent with the Center's mission and targeted to important needs for improving schools. Thus, from an operations perspective, our strategy from the beginning was to match talent and resources at the consortium sites with the tasks necessary to fulfill the Center's research goals.

Our planning efforts were consequently focused on using literature and research findings, the results of our needs assessments, and the advice of consultants and practitioners to select needs and strategies that would provide direction for the Center for years to come. One direction was a program of research on uses of tests.

Program on Uses of Tests

There are distinctly different uses of tests in education associated with particular groups of users. Teachers use formal and informal tests for instructional decision making. School administrators use them for curriculum review and resource apportionment decisions. Counselors use them for advising students and teachers. Policy makers in education use them to monitor education systems and to influence educational programs. Citizens use them to gauge the effectiveness of schools. Yet there is very little research on these groups' information needs or their motives for testing and the extent to which their needs or intents are, or can be, met by current or future testing practices.

This program of research would examine the use of tests for the different purposes listed above, asking whether the needs or intents of the client groups are being adequately met by existing testing practices, and discovering ways to improve testing practices so that they better serve consumer groups.

A distinction has been made between internal and external testing programs depending on locus of control--within or outside the school district. School building or district-wide achievement-monitoring testing programs would be internal, while state-mandated or legislated programs and Chapter 1 testing would be external. Another distinction may be made between tests used to inform users and those used as administrative devices (e.g., to control award of high school diplomas or to determine grade promotion). This program of research would seek to investigate the uses and impacts of both internal and external testing intended either for informing consumers or for administrative intervention.

Initially, researchers at the R & D Center could examine the use of tests in policy. During the last five years, efforts to reform education, particularly at the state level, have increasingly employed tests and test results in various ways, and this use of tests in the policy sphere is a growing trend. For example, a 50-state survey of reform measures conducted by **Education Week** (February 6, 1985) found the following: 29 states require competency tests for students, and 10 other states have such a requirement under consideration; 8 states employ a promotional "gate" test, while 3 others are considering such a mandate; finally, 37 states have some sort of state assessment program, and 6 additional states have such a program under consideration. This growing use of tests in the policy sphere cannot help but impact on teachers and students, as well as on more traditional testing programs, evaluations of students and standards of educational excellence. We felt therefore, that it is imperative that the NIE Center document the impacts--both positive and negative, the costs and the benefits--associated with these external testing programs. Further, we felt that there is a pressing need to develop practical strategies and techniques that state departments of education and local school systems can use to evaluate these programs and to make better use of the information they can provide.

Another line of inquiry within this program of research would be an examination of testing and standard setting practices in Australia, England, Ireland, Germany, the Netherlands, Sweden, and Japan. There is speculation that the testing practices in these countries have much to offer to reform efforts in the U.S. (Madaus & Greaney, 1985), yet there has never been a systematic and thorough study of the strengths and weaknesses, contextual impacts, and utility of alternative practices so that testing deficiencies in this country can be matched to strengths of the practices of other countries. In order to build on the experiences of others, we must know what they have been and then evaluate their potential for addressing our needs. This project should examine and describe current practices in the selected countries, compile information about their strengths and weaknesses, investigate side effects and contextual idiosyncrasies of each, and then develop recommendations for testing reform in this country.

This program of research on uses of tests was directly linked to the mission of the NIE R & D Center. It focused on improving the use of formal and informal tests in both internally and externally controlled testing programs, expanding our knowledge of both positive and negative impacts of testing, and developing products that educators can use for school improvement. It can have immediate and tangible payoff for students who may be adversely affected by external testing policies by informing and enlightening the policy-shaping community. It can expand equality of educational opportunity by making important information about students more accessible to educators and less open to misinterpretations. It will closely examine existing practices and lead to the development of new methods and uses of testing.

Priority projects that were identified included:

- Use of Tests in Policy
- Use of Tests in Schools
- Visiting Scholars and Practitioners Project
- School Partnerships to Integrate and Test the Center's Products
- Studies of Testing and Standard Setting Practices in Other Countries

The research on the use of tests in policy was chosen after a careful review of related literature. During the last five years, efforts to reform education, particularly at the state level, have increasingly employed tests and test results in various ways. First, tests have been used to inform policy makers about the state of education, and second, they have been used as administrative mechanisms to drive policy. The latter is accomplished by attaching important rewards or sanctions to test performance: high school diploma, compensatory funding, merit pay, district certification, etc. The use of tests in the policy sphere is a growing trend and cannot help but impact on teachers and students, as well as on more traditional testing programs, evaluations of students, and standards of educational excellence. Therefore, we proposed to document the impacts--both positive and negative, the costs and the benefits--associated with these external testing programs. Further, we saw a pressing need to develop practical strategies and techniques that state departments of education and local school systems can use to evaluate these programs and to make better use of the information they can provide. This work should help to minimize the potential negative outcomes or abuses that can result from such programs.

A consideration of the possible effects at the school level of using standardized tests suggested that one might expect effects on school organization and on a number of school practices. Two relatively recent studies of the impact of more traditional testing programs indicated that such programs have little effect on school level organization or administrative decisions. Sproull and Zubrow (1981), after an intensive, small-scale study in Pennsylvania, concluded that test results from traditional school district testing programs were not very important to central office administrators and that administrators are not major users of test information. In an experimental study of the effects of introducing standardized testing in the schools of the Republic of Ireland, Kellaghan, Madaus, and Airasian (1982) found that school principals, when questioned about various aspects of school organization and practice, indicated that, at the administrative and institutional levels, the overall impact of a standardized testing program was slight. The findings extended to a wide variety of functions including: admission to school, the content of school report cards, streaming practices, provision for remediation and referral, communication practices, retention in grade, and the curriculum.

In interpreting the negative findings cited above, one must keep in mind that they speak only to the effects of information from traditional school district testing programs on institutional organization and practice. They were carried out before the advent of state-mandated testing programs aimed at reforming education, and before recent efforts by school superintendents to use test information to drive instruction (sometimes referred to as measurement-driven instruction) or before the use of tests to continuously monitor student achievement (sometimes referred to as continuous achievement monitoring, or CAM programs).

The large-scale research that is available on teacher-level effects of standardized testing is based on surveys of teachers, most of whom had had considerable experience with such testing (Goslin, 1967; Beck & Stetz, 1979; Salmon-Cox, 1981; Kellaghan et al.,1982). These surveys show that standardized tests, while viewed favorably by teachers, were not of great relevance in their work. On the other hand, logic suggests that when test results carry with them important consequences for pupils, teachers, or schools (as is the

case with many externally imposed testing programs), then the perception of their relevance might be quite different. Teacher perceptions of test relevance might also be quite different when the test results are part of a local measurement-driven instruction program or a CAM program.

It seems reasonable to assume that standardized test score information has its most serious impact on the student. Thus, it was not surprising to find that most research on the effects of standardized testing has been concerned with effects on pupils.

Goslin (1963) suggested two levels at which test information might affect students. The first level is the direct impact of providing students with additional information about his/her own abilities in the form of test scores. The second level of effects on students are those that result from communicating test results to other people who in turn take actions that impact on the student(s).

Bloom (1969) has argued that if the tests are understood and utilized properly by students and teachers, they can do much to enhance a student's learning as well as his/her self-concept. On the other hand, it is conceivable that learning the results from a test might adversely affect an individual's self-concept, level of aspiration, or educational plans. Empirical evidence relating to the impact of providing students with test information in noncognitive areas is surprisingly scant.

Part of the reason that research in this area is so sparse is the complexity of investigating the issue. Important distinctions have to be made between the pupil's age (test results often are not directly communicated to young pupils but are to secondary students); the kind or amount of information provided (norm- or criterion-referenced information, achievement or ability information); and the type of testing program involved (external test with important sanctions associated with the results or traditional school-based testing programs). The measurement of the students' self-concept is also no easy task, as self-concept is not a unitary trait.

While there are no hard data available, there was considerable anecdotal evidence presented at the 1981 NIE-sponsored hearings on minimum competency testing that many students who failed a graduation test the first time dropped out of school never taking the test again. Whether or not the decision to drop out is directly or indirectly related to failing the graduation test is unknown (transcriptions and videotapes of these hearings are available from NIE).

Providing test information to students may also directly impact on their academic performance. This is the belief behind many measurement-driven instruction, CAM, and mastery learning programs. Further, this is an argument often made to justify state-level testing programs, particularly those directly linked to promotion or graduation decisions.

Popham, Kruse, Rankin, Sandifer, and Williams (1985) report that student test scores have risen dramatically in Texas, Detroit, South Carolina, and Maryland. In all of these locations, measurement was perceived as a catalyst to improve instruction. In addition, a number of people have pointed to a sharp decrease in the numbers of students failing minimum competency gradu-

ation tests as evidence of the program's success. However, alternative explanations for these gains have not been sufficiently explored. They may be due solely to teaching to the test. They may not generalize to other measures of the same construct and in fact may change the original construct the test was designed to measure. In all of the programs cited above, students presumably were made explicitly aware of their performance. Nonetheless, it is difficult to ascribe the gains in test scores solely to this fact, since the teachers' instruction presumably also changed to come in line with whatever the tests were measuring.

LeMahieu's (1984) evaluation of a CAM program in the Pittsburgh Public Schools shed some light on how these measurement-driven instructional programs work. His results indicated that the program had generally positive effects on students' achievement as measured by test scores. He found that the CAM program clearly focused the attention of students and teachers on the skills to be measured, and that this largely accounts for the improvement in achievement. However, LeMahieu pointed out that this focusing phenomenon also raises the following concerns:

1. the routinization of instruction by some teachers who may adopt the objectives of the monitoring program as the sole content of instruction in that domain

2. a loss of residual learning outside of the CAM content

3. as additional areas of the curriculum are added to the CAM program, they might begin to compete for an extremely important and limited resource--instructional time. In fact, Pittsburgh teachers reported that they took the time for supplemental instruction in math (the area covered by CAM) away from other subjects. LeMahieu suggests that these difficulties can be overcome by careful planning and wise management but that these dangers are real and ever present.

Three different theoretical frameworks guided our approach to this research, namely aspects of organizational theory, information theory, and the multiple methods approach to research. We cannot elaborate here in any detail on how each of these perspectives guided the proposed research. Hence, with respect to each we simply give one prime example of how each guided our thinking.

Sociologists of educational organizations have in recent years characterized schools as "loosely coupled" organizations in which units such as schools within districts and teachers within schools have a fair amount of autonomy in carrying out their duties. A related concept is that of teachers as "street level bureaucrats;" that is, government employees having explicit responsibility for carrying out state policies, but having much autonomy in their day-to-day work and, burdened by many demands on their time and limited resources, inevitably having to engage in many accommodations in order to carry out their duties. These ideas provide important perspectives on schools as compared with more tightly structured organizations (such as many business organizations), but obscure the fact that schools are in fact part of an educational system structured in highly hierarchical fashion, with states having constitutional authority for education, school districts typically being given considerable leeway in implementing educational policy, and

teachers within schools having much autonomy in carrying out their day-to-day responsibilities.

These considerations led us, in looking at the impact of testing, to two important general points. One is that one must examine each of these levels of educational organization in order to thoroughly understand the effects of testing. Thus, various of our hypotheses are aimed at each of these levels. The other is that because schools are loosely coupled organizations, with much room for accommodation in much of the day-to-day routine of instruction, to understand the role of testing we need to delve beneath official policies in order to see how testing affects the accommodations that teachers and students inevitably must make in meeting the demands of complex social organizations.

The key idea drawn from information theory with respect to the proposed research program was simply that one cannot judge the value of a particular piece of information--or kind of information--in a vacuum. One must also look at the "signal to noise" ratio; that is, the possible variety of competing or confirming information in addition to test information that may bear on a particular issue. The importance of this point was aptly demonstrated in Raudenbush's (1984) meta-analysis of teacher expectancy studies which clearly showed that the expectancy effects of test information were greatest when teachers had little previous experience with students, and hence a relatively small store of previous information on students which might influence their opinions. Though this point may seem obvious, it represented a perspective which has not much informed previous research on the effects of testing on curriculum and instruction. Thus, in looking at effects of testing on these broader aspects of educational organizations, we must be alert to alternative sources of information and influence on curriculum and public opinion regarding schooling. It also illustrated the importance of the distinction we drew between using tests as administrative devices versus using them to inform policy. When tests are used as administrative devices, the signal-to-noise ratio of test information obviously is raised.

The third key perspective informing our research was the multiple methods approach to research itself. It is well established in social research that the particular methods of inquiry one uses affects both what one looks for and what one sees. In psychological research, Campbell and Fiske (1959) are known for proposing their multitrait-multimethod approach to construct validation, but their general perspective is relevant to other forms of social research, including research on the effects of testing. Therefore, in looking at the effects of testing, we need to employ multiple methods of inquiry, looking not just at patterns of test score performance over time, but also at survey evidence, informed opinion gathered through interviews, and quasi-experimental evidence, as discussed in the procedures section of our full proposal. When results from different methods of inquiry converge, we can have confidence that results are not merely artifacts of one particular method of inquiry employed.

The primary research approaches that would be employed in the study of policy uses of tests were proposed to be surveys, interviews, and quasi-experiments.

The program of research on uses of tests also included a proposal to study the uses of tests in schools. There have been several naturalistic

studies of test and evaluation use in schools completed over the past ten years (Alkin, Daillak, & White, 1979; Kennedy, Apling, & Neumann, 1980; Rudman et al., 1980; Radwin, 1981; Salmon-Cox, 1981; Sproull & Zubrow, 1981; King & Pechman, 1982). Based on these studies, it is safe to conclude that neither testing nor evaluation has been well integrated into the everyday practices of classroom teachers and school administrators. That is, the potential of testing and evaluation for improving student growth and development has not come anywhere near realization.

When confronted with the reality of the minimal role of testing and evaluation in schooling, one response has been to abandon them in favor of less systematic, less formal means of generating information for educational decision making. Lortie (1975) and Kennedy, Apling, and Neumann (1980) have documented occurrences of this response in schools. A side benefit of such a response is the elimination of misuses of testing and evaluation that have been reported in recent years (Holmen & Doctor, 1972; Brickell, 1976; House et al., 1978; Madaus, Airasian, & Kellaghan, 1980).

To throw out testing and evaluation altogether, however, is akin to throwing out the baby with the bath water. Benefits that may be derived from information generated through testing and evaluation have been widely discussed and accepted (Stufflebeam, et al., 1971; Kellaghan, Madaus, & Airasian, 1982; Haney, 1984). These include: (1) identification of student needs; (2) guidance for selecting among known alternatives in instruction; (3) reductions in the influence of prejudice and impressions in making decisions affecting students; and (4) justification of expenditures of public funds. Without results of testing and evaluation programs available, school administrators have been found to engage regularly in personnel and program decisions (such as making assignments, planning in-service programs, setting goals for individual schools, making budget allocations, and selecting program designs and materials) based on limited and often subjective information.

Armed with this knowledge and an understanding about how testing and evaluation are perceived by school staffs, researchers need to ask how testing and evaluation practices can be tailored to fit the typical information needs of school teachers and administrators to help them make better decisions and hence to improve the quality of instruction offered. Can the routines just described be made easier for school staff members, while also being made more accurate, systematic, fair, and comprehensive? "Is there a better way to do it?" is a question that may be asked of each routine activity. Research on that question, keeping sight of the need to make life easier for school staff, is the challenge to be addressed by this research project.

The purpose of this project was to gain a better understanding of the information needs of teachers, principals, and other school professionals and how these needs are best served by improved formal and informal testing methods, reporting and test interpretation techniques, and school-based microcomputer systems.

Specifically, the objectives of this project were to conduct:

1. descriptive case studies of student testing and evaluation practices in two schools and two school districts in the greater Kalamazoo, Michigan region;

2. psychophysical studies of how teachers, principals, school district administrators, counselors, and school board members interpret and use simulated test results;

3. a feasibility study of transporting a school-based microcomputer student information system from the school where it has been developed (the M. L. King School in Pittsburgh, Pennsylvania) to another school (near Kalamazoo, Michigan) which has no prior experience with the system.

Based on the results of the final year of this project, plans for studies of information needs of teachers, principals, and other consumer groups for testing in education and how these needs can best be served, could be prepared for use in subsequent years of the project. In particular, practical support materials to aid in the reporting and interpretation of test results could be designed and pilot-tested. Also, experience gained from the feasibility study of school-based information systems will be used to design and test student information systems in a variety of school settings.

A second line of inquiry in this project was to plan to conduct psychophysical studies of how teachers, principals, school district administrators, counselors, and school board members interpret and use simulated test results. We would develop and present hypothetical test results to samples of teachers, principals, school district administrators, counselors, and school board members drawn from at least ten different school districts and ask them to perform a variety of different classification tasks based on the test results. The stimulus test results would be systematically varied on at least three dimensions (namely, the title given to the test, the scale in terms of which results are reported, and the precision with which results are reported).

Analyses of the ways in which people perform the classification tasks would provide evidence on which to base interpretations and reporting of test results in different ways.

A third line of inquiry was to conduct a feasibility study of transporting a school-based microcomputer student information system from the school where it has been developed to another school which has no prior experience with the system. Staff at the Learning R & D Center in Pittsburgh have developed a prototype system for making necessary student information available to school personnel. The feasibility of transporting this system to another school in the greater Kalamazoo, Michigan area could be investigated to assess its potential for responding to the real information needs of school staff and the transferability of this technology.

R & D Center staff would arrange with several school district superintendents and building principals to conduct interviews with their school staff about the information they currently have and use. Classroom teachers and building principals would investigate, through panel discussions and faculty meetings, how such a system could work for them and how it would have to be adapted to make it useful and workable. Reports from each of the schools will be used to describe the feasibility of using a microcomputer to better integrate testing into instruction and school improvement.

Interviews and observations would be conducted. Inservice workshops would be held. And use of the system would then be discussed. A report of the

feasibility of using such a system to better serve the testing needs of local educators, with recommendations for further research and development would then be prepared.

The program of research on uses of tests also included a proposal for a visiting scholars and practitioners project. We believe for the Center on Student Testing, Evaluation, and Standards to be effective, it must (1) ensure that its programs are responsive to the needs of schools and (2) integrate the products of its programs and projects into systems that will work in schools. The purpose of this project is to collaborate with practitioners and scholars to: a) assess needs in schools related to testing, evaluation, and standard setting; b) evaluate the relevance and practicality of the Center's contributions from individual research and development projects; c) plan the Center's agenda of installation and demonstration projects; and d) set up one prototype school-Center partnership project aimed at helping a school to improve its collection and use of testing and evaluation.

The operational framework for this project is a two-year colloquium. The participants would include school personnel from the Kalamazoo area, Center personnel, and visiting scholars and practitioners. The participants would collect needs data through surveys and a study of a selected pilot school, study the reports of the Center's projects, review relevant related research, assess the possibilities of combining the projected contributions of the Center's other projects for use in schools, engage in collaborative planning with one school to set up a school-partnership project, and document what is learned through the colloquium experience. In a very real sense this group would help the Center Director set the stage for the Center's development projects to be conducted in 1988 and 1989. Those projects are projected to include several school-Center partnership projects.

The benefits from this project would be of five major types. The Center could gain a more concrete view of the kinds of assistance that schools need in the areas of testing, evaluation, and standards. Feedback from the project could be used to assess and improve the relevance and practicality of the Center's programs. One school would be assisted to assess, synthesize, and operationalize the contributions from the Center. A model plan for a school-university partnership project would be provided. An agenda of developmental school-Center partnership projects would be developed. The consulting practitioners and scholars would produce publications, based on their work in this project, for use by both researchers and practitioners. And, the participants in the colloquium would be provided a rich learning experience.

Common criticisms of research in education are that it has had little influence on practice and that the timelag between invention and adoption of educational innovations is long (for example, see Huling, Richardson, & Hord, 1983). Among the reasons given for this less than optimum impact are:

- the failure to involve practitioners and the schools in the planning and implementation of research and development
- the use of highly structured methods, such as experimental design, that do not accommodate the realities of the classroom, school, and school district
- the isolation and remoteness of university research programs from the setting of classrooms and schools

- failure to recognize a focused perspective of the schools' problems and their possible solutions
- development of programs in isolation without engineering them to fit the context of the school or school district

The members of our Center have had a long history of working with schools and school districts. While we have recognized the need for developing discrete projects with reachable goals, we also recognized the potential hazard of developing procedures and systems that fit the development context but not that of other school situations. This project would provide a main strategy for pulling together the Center's other projects, integrating them, and making them responsive to the problems of practitioners and the schools.

The overall aim of this project is to develop and implement a strategy for maximizing the contributions of Center projects through collaborative relationships among scholars, practitioners, participating schools, and Center personnel. To accomplish this goal, the primary subgoals of this project are to involve practitioners, scholars, and staffs of the Center and a selected school to:

1. identify needs of school personnel that relate to the mission of the Center;

2. evaluate the potential contributions of the Center's research and development projects to address the assessed needs of selected schools;

3. create an agenda of development projects that reflects the needs of selected schools and provide for integrating the contributions of the Center's other projects into systems that can be tested in schools;

4. develop an agenda of school-Center projects to be implemented in years three and four of the Center's operation;

5. set up a model for school-Center partnerships in the mission area of the Center.

A secondary goal of this project would be to capitalize on the expertise of the participating practitioners and scholars to produce documents of general interest to the educational community and of interest to specific audiences.

Accomplishing the above goals would greatly increase the ability of practitioners to influence the Center and increase the probable impact of the Center's research and development on school practice.

Program on New Technologies in Testing.

A second direction that our literature reviews, needs assessments, and interviews took us was toward research on new technologies in testing. Current reform efforts across the United States are emphasizing that education must be more accountable. Measures of student achievement are looked upon as the ultimate indicator of the success or failure of the schooling process. This increased attention on student testing is presenting a serious challenge to the education profession as to how new systems and techniques can be developed to assess student achievement.

The advent of microcomputers with memory capacity and operating speeds to rival mainframe computers of 15 years ago has opened the possibility for major advancements in how we assess students. Increased capabilities in the field of communications also present possibilities for improving assessment strategies. Consistent with the proposed mission of this center, this program would identify the best strategies within testing and combine these with the latest innovations of technology.

Merging technological advancements with the best in testing strategies can be achieved through a collaborative effort, one that includes participation from universities, industry, and local school districts. This consortium, with the participation of industry leaders, provided a natural base for a major effort in developing and demonstrating new systems and techniques in student assessment.

New systems would be developed on a project basis with priority need areas being identified from input provided by school personnel and measurement experts. Each project would be developed in a systematic manner with detailed monitoring and feedback. Student data for each project would be collected from diverse school systems, including Dallas, Kalamazoo, and Pittsburgh. The projects would be enhanced by carefully designed evaluations from both internal and external evaluators. The control provided through pilot testing in the participating school districts would assure products that are both state-of-the-art and practical.

Priority projects that were identified included:

- Adaptive Testing - computer-driven testing of students on items based on an existing curriculum. Each student would be tested on items adapted to his/her level of capability or level of instruction.

- Data Based Decision Making - Strategies would be developed and researched to find the most effective ways to display and interpret test data to enhance decision making.

- Uses of Tests with Bilingual Populations - In depth work with teachers to determine and respond to the unique testing problems of bilingual students.

- Diagnostic Testing and Instructional Management - Using adaptive testing strategies, methods would be researched to optimally impact the instructional process. Students in such diverse populations as special education and talented/gifted could be tested for accurate placement within the instructional process and monitored for achievement progress.

Future research projects would be developed to focus on problems of testing with mildly handicapped populations.

Current research in testing, item bank theory, and adaptive testing indicates a need for a well-articulated, pragmatic study of adaptive testing of student achievement and mastery of curriculum goals. Our proposal described a research project that uses current curricula and item sets in mathematics at the elementary level to create an integrated, microcomputer-based system for

adaptive testing of student achievement. The major components of the system included (1) a calibrated and instructionally indexed item bank for testing progress in the curriculum, and (2) a user-friendly, integrated software system that allowed for:

a. the selection of tests and test items through the bank indexing and calibration system

b. the production of placement and diagnostic information for use by teachers in assessing student progress

c. the production of student-performance based data for instructional planning and management

d. summative assessment of student progress in the curriculum

e. two-way communication with mainframe computer systems

The research outcomes of the proposal included data relative to the problems of:

1. psychometric bases of item bank indexing and calibration systems

2. item bank indexing and calibration for use in assessing student achievement of curricular material

3. measurement characteristics of tests constructed from indexed items and the robustness of these measures in pragmatic applications

4. the use of adaptive testing data in conducting and assessing standard setting

5. instructional refinement through adaptive testing

6. the use of diagnostic test information in formulating strategies for instructional management

With the technological breakthroughs in microcomputer design in the past five years, computerized adaptive testing has become a practical, as well as an affordable, concept. Many discussions have taken place regarding the potential for the use of a microcomputer-based adaptive testing system (ATS) in conducting student placement, diagnosing student problems, and assessing student achievement. The measurement issues have been discussed, and the technological potential has been assessed. Most of the discussion and research, however, has been hampered by the lack of an actual system to use in testing the hypotheses and the robustness of the assumptions.

Beyond the need for a carefully constructed ATS, a practical void has existed in the realm of linking an ATS and its output to classroom instruction and instructional management. Our proposal was designed to fill that gap. The proposal was to create an ATS based on current research using existing curricula and existing test items in elementary school mathematics. This system would use an integrated software management system to create and administer tests, provide instructional feedback relative to student progress,

provide instructional management information to teachers, provide data for instructional standard setting, and communicate with mainframe data bases.

The research goals of the project would be to use the constructed ATS to provide information relative to the problems of:

1. the psychometric basis of item bank indexing and calibration

2. item bank indexing and calibration for use in assessing student achievement of curricular material

3. measurement characteristics of tests constructed from indexed item banks and the robustness of these measures in pragmatic applications

4. the use of adaptive testing data in conducting and assessing standard setting

5. the refinement and improvement of instructional management through the use of information through the ATS

Currently, none of these areas have been investigated in the context of a functioning ATS.

Related research on ATS use has been focused in three areas. The first has been concerned with the issues in item banking and the corresponding use of item response theory. The second area has been the technical characteristics of ATS-generated tests. Third has been research related to the potential for using a microcomputer-based ATS to improve student assessment, achievement, and instructional management.

Millman and Arter (1984) and Wright and Bell (1984) present useful overviews of item banking that discuss the broad issues associated with its applications. The most useful part of the Millman and Arter discussion is an outline of the major issues involved in item banking and issues of concern for those attempting to create or evaluate an item bank. Wright and Bell discuss the mathematical and psychometric foundations of item banking in addition to discussing its potential. Both studies suggest the potential for using item banks in student assessment. They also say such banks can provide information for teachers and for curricular improvement. While both studies discuss the potential for curricular improvement, however, neither cites any actual attempts at implementation or offers any insights into how the linking between an ATS and instructional management would be accomplished.

Green, Bock, Humphreys, Linn, and Reckase (1984) provide a discussion of the characteristics of computerized ATS systems. They address item response theory, dimensionality, reliability and measurement error, and validity. They conclude that adequate procedures exist for assessing the properties of such systems, but that a great deal needs to be learned about them and their potential. Kreitzberg and Jones (1980) presented the results of an empirical study of a minicomputer-based ATS for a test of verbal ability. They point out that such systems are pragmatically feasible, that a need exists for trials of these systems in the field to assess student achievement rather than aptitude, and that a need exists to investigate microcomputers as the delivery media for these systems.

Hambleton and Swaminathan (1985) suggested three important benefits of using item banking. First, test developers will easily be able to build tests to measure objectives of interest. Second, test developers will be able to use item banks to produce tests with the desired number of test items per objective. Third, when item banks contain content-valid and technically sound items, test quality will usually be better than what test developers could produce themselves.

Hambleton, Anderson, and Murray (1983) indicated great potential for the use of microcomputers in classroom testing. They report possible uses of ATS systems for improving instruction, more accurately assessing student ability, providing immediate student feedback, etc.

The foregoing examples illustrate the potential of item banking and ATS for improving classroom instruction. However, before this potential can be completely realized, certain needs must be met. For example, several questions about the theory underlying ATS systems and their properties must be answered. In addition, fully operational versions of such systems in classrooms settings must be created so that the explicit links between implementing an ATS and using the results to improve instructional management can be determined and tested in a natural setting.

This proposed research had direct applicability to the Center's mission and provides a vehicle for the investigation of other problems germane to the Center. It also crossed into the research domains of other proposed centers and thus provides a means for collaborative research. It was first and foremost directly applicable to the Center's objective of investigating new technology and its applicability in the classroom to improve measurement and learning. The proposal provided for a system that incorporates state-of-the-art technology with the cumulative research in ATS theory to give a curriculum-based system with research applications in standard setting, student testing, evaluation, and psychometric theory. It also offered the opportunity for researchers to gather a wealth of information from field tests in actual school settings and establish item banks in content areas and grades where improved testing is greatly needed. This was particularly relevant given the current interest in elementary mathematics instruction.

Finally, this project offered a link with the investigation of instruction and the coordination of instruction and testing by its ties to existing curricula and test items in use in school systems. These linkages are also established by the indexing of the item bank according to the instructional objectives of the curricula. In line with the Center's mandate to collaborate with the other proposed centers, this project offered ideal ways for linking with those centers that deal with instruction and instructional improvement and centers dealing with teacher education.

Program On Standard Setting.

The third direction that our review of literature, needs assessments, and discussions took us was toward a program of research on standard-setting. This program was keyed to understanding the nature of educational standards held by our citizenry, developing methods that are capable of accurately reflecting societal expectations, and examining the consequences of standards on educational processes and outcomes. Today there is considerable public concern about the standards of our schools. Standards evolve from societal needs and

are then shaped by what is taught and eventually by the manner in which we judge what has been learned, e.g., through testing. Clearly, standards do not exist in a vacuum. They affect and are affected by what is taught and what is found by way of testing or evaluation. Given the central role that standards play, there is a need for systematic inquiry on standards so that the complementary forces of testing and evaluation and of teaching and administration contribute to improved school practices.

A major line of inquiry on standards that we planned to undertake was geared to understanding the process by which standards evolve. Evidence of educational standards can be seen from classroom grading practices to state requirements for high school graduation. Yet little is known of the social, cultural, political, and economic factors that contribute to the establishment of standards or the role played by the various stakeholder groups when standards are set or accepted and used. Further, the audience education serves is diverse, and studies are needed to examine the relationship of standards to the values and needs of distinct populations. Utilizing case studies at the level of teachers and the local schools, and at district and state levels, could lead to an understanding of how standards are formed. Investigations were also planned to address standards and the process of their formulation across what are judged to be effective and noneffective schools.

A second area for investigation addressed the methodology of standard setting. While the recent past has witnessed a great many reported studies of the characteristics of standard-setting techniques, research to date has been narrowly focused. There is a need for research on issues such as the effect of directions on resulting standards, clear definition of the components required if equitable and acceptable standards are to be set, examination of the appropriateness of fixed standards for different populations, exploration of alternate methodologies for standard setting, and issues of test instructional validity tied to standard setting on tests with different purposes (e.g., CRTs vs. NRTs).

Beyond these properties of methods, technical questions remain regarding such issues as test length and composition, appropriateness of various standard errors for deriving accurate cut points, the utility of multiple thresholds, and the extent to which test dimensionality, defined by either content domain or test structure factors, confounds standards. If fair and equitable standards are to be established, these are among the many issues requiring attention.

Research was also planned to address the question of the consequences of standards on educational practices and outcomes. While there has been considerable speculation and debate, this area has not been systematically studied. A few of the questions to be addressed include: Do standards prescribed as minimal become maximal? What are the consequences of teacher/test classification discrepancies? How can test result information be best reported to facilitate use? What form of information is most useful for different audiences? What are the shifts in standards over time? In what ways do standards affect school processes?

While questions relating to standards are to be addressed at a number of levels, our focus was largely at the local/community level, for it is at this level that test and evaluation results must be responsive if school learning is to be maximized.

Priority projects that were identified included:

- The Process of Standard Setting in Effective and Ineffective Schools
- Study of Methods and Techniques for Setting Standards
- Competency Testing and its Impact on Educational Standards
- Methods and Impact of Reporting Test Results to State and Local agencies

That American schools have failed to set and maintain proper standards for what is learned and how learning is judged is a proposition that enjoys virtually universal national consensus. General reports of the schools' failings are common (see, for example, Boyer, 1983; Goodlad, 1983; National Research Council, 1977; National Commission on Excellence in Education, 1983; Sizer, 1984), as are reports dealing with specific curricular areas (e.g., College Entrance Examination Board, 1977; Chall, 1977). Proposals for remedying this problem have also appeared (e.g., Adler, 1982; Resnick & Resnick, 1985).

Despite this surfeit of information, the question of how standards actually evolve at the local school district level remains for the most part unanswered. If a school district determined to respond to the national cry for accountability and establish new and presumably higher standards, what are the ways it should go about that task? Even supposing that a district had complete and accurate information on what little the research has shown on these matters, how could it productively use that information?

There are multiple factors involved in reaching appropriate decisions and many of these factors are in conflict. For example, schools seem to be expected to share a common national set of standards; indeed, much of the literature mentioned in the preceding paragraph implies that national standards are imperative. Yet politically, schools are under local control. How can a standard-setting process take account of this bifurcation in power? The proposal that there be national standards presupposes that there is a viable mode for determining what they should be. But can that be so? The nation's experience with, and our own understanding of, cultural and value diversity suggest otherwise. "National standards" also suggest some common core of knowledge and skills that all should possess; however, for a century it has been an article of faith among educators that good teaching caters to and capitalizes upon individual differences. The national desire to maintain equal opportunity for all can be taken to imply that everyone should be provided a college preparatory curriculum so as not to deny the student the option of college. On the other hand, tracking practices and the demands for adequate vocational education seem to preclude such opportunity for many. Standards are often said to be set by textbook publishers, who may have their own reasons for determining book content. To the extent that this is true, what flexibility adheres to the local district, especially in states that regulate textbook adoption?

It seems clear that standard setting is not simply a matter of deciding what to do and doing it or deciding to do it better and doing that. The standard-setting process historically seems to have been less a matter of making rational, research-based decisions than making compromises, adaptations, and accommodations with a variety of standard-setting forces that seem

almost to have a life of their own, one well beyond the capability of the typical local district to control. And it seems likely that efforts to alter (hopefully to raise) standards will not be successful unless a great deal more is understood about how that process occurs "naturally." It is even questionable whether standard setting has been a process in which local districts have been consciously involved. They may simply have been forced into compliance with standards shaped by outside forces.

The purpose of one project that we proposed was to provide insights into the question of how standards actually evolve at the local school district level. It seemed clear that if effective approaches are to be taken toward setting new standards, it is imperative to know something about how the process occurs in a real-life setting.

The proposed study has three major objectives:

1. To study in situ several school districts with a view to describing how standard setting occurs in each. This objective will result in one or more (probably two) case studies that portray the standard-setting process (or lack of it) as reconstructed by "insiders," that is, an **emic** view, as opposed to the **etic** (or outsider's) view.

2. To test, as part of the process of carrying out the first objective, assertions currently found in the literature about the standard-setting process, as well as current recommendations for its improvement. Can evidence be uncovered to suggest that existing research descriptions, which are essentially generalizations putatively "true" of a statistically average district, hold true in particular and concrete settings?

3. To make recommendations of two different sorts as an output:

    a. Considering the case studies carried out in this project how can other case studies be mounted, perhaps purposefully selected to contrast with those already completed along dimensions of factors which the initial studies suggest are important? What suggestions emerge from the present study to improve the methodology to be employed?

    b. Considering what is suggested by the case studies about the degree to which factors described in conventional studies work themselves out at the local level, how can future scientific studies of standard setting be more adequately grounded? How can the methodology be altered to expose the unique adaptations and accommodations to local conditions?

A second project on standard-setting was proposed to address a very different gap in the literature, using a very different research strategy. A review by Berk (1985) revealed that no fewer than 30 methods have surfaced and been used within the recent past to set standards for test performance. For the interested reader a description of the three most commonly used methods (Angoff, Ebel, and Contrasting Groups) is provided below.

The Angoff, Ebel, and Contrasting Groups Procedures are based on expert judges' assessments with respect to the expected performance of students. The methods differ in terms of the specific factors rated. For these methods, both the judgments made and the standards derived are independent of the actual performance of students on the tests.

Angoff method: For each test item, raters estimate the probability (on a scale of 0 to 100) that a minimally competent student will know the correct answer to the item (Angoff, 1971). In essence, the judges estimate the difficulty level of an item, referencing a hypothetical group of individuals that would be judged minimally competent. To obtain the overall standard, probabilities assigned by a judge are summed, then averaged over judges to yield the passing score. This standard represents the estimated mean total score for a group of minimally competent individuals.

Ebel method: Judges make three judgments (Ebel, 1979). First, judges rate each test item on two separate dimensions: level of difficulty (easy, medium, or hard), and degree of relevance (essential, important, acceptable, or questionable). Then a judge indicates the proportion of items to be answered correctly for each difficulty and relevance configuration, e.g., easy items that are important. To derive the standard, each item is assigned to its appropriate cell based on the judges' ratings. The percentage passing judgment for a cell is then multiplied by the number of items in that cell, and these products are summed over all cells to obtain the passing score for a judge. Passing scores are then averaged over judges to obtain the passing score for the test.

Contrasting Groups method: A teacher classifies a student into one of two groups, Competent or Non-Competent, relative to the content being assessed (Livingston & Zeicky, 1983). Based upon these group membership classifications, and the actual test scores of these students, a standard is derived, using statistical likelihood-ratio procedures which minimize the probability of misclassification of students into groups. There are several variants in the specific statistical procedures available. Choice of a procedure is dependent upon the population distribution shapes and relative variances of the two groups' test scores.

The methodology of standard setting has become the most researched topic in the criterion-referenced measurement literature. A summary of the findings from this literature follows:

1. Two general classes of procedures for setting test standards have emerged and now hold center stage: judgmental item review-based methods, referred to as "Continuum Models" and empirical performance-based methods, referred to as "State Models" (Jaeger, 1976; Meskauskas, 1976).

   Within the judgmental item review class are those methods such as the Angoff, modified Angoff, Ebel, Jaeger (1978), and Nedelsky (1954) procedures. The use of such methods requires that raters, those persons charged with determining the performance standard, examine the actual test items and offer opinions as to how students are likely to perform and/or judge the item's relevance. The empirical

performance methods, e.g., Contrasting Groups, Borderline Group, and Criterion Groups, derive a standard by examining statistically how persons tested actually perform in relation to how they were expected to perform given teacher nomination.

2. Different methods, regardless of their class, produce different standards (Andrew & Hecht, 1976; Poggio, Glasnapp, & Eros, 1981, 1982, 1983).

   This finding has been reported consistently. Insofar as different methods ask different criterion questions of the standard setters, this is not an altogether surprising result.

3. The level of the standard produced by a method in comparison to standards resulting from other methods can be fairly well predicted (Berk, 1984; Kottler, 1980; Poggio & Glasnapp, 1981, 1982; Skakun & Kling, 1980).

   The literature is consistent on this point and has demonstrated that regardless of content tested, grade of students being tested, affiliation of the group setting the standard, or the nature of the decision to be made, a method can be expected to result in a standard systematically different from that of other methods. The reasons for such trends are not altogether clear at this time.

4. Regardless of the method used, the standard derived is susceptible to both measurement and sampling error (Berk, 1976; Hambleton, 1984; Millman, 1973; Poggio, 1984).

   All methods require judgments to be made, and the process of judging is fallible. Further, it has been shown that the relationship of judges to group membership (e.g., teachers, administrators, policy makers, parents) is related to the level of the resulting standard.

5. In most applications performance is measured on a continuous scale, and to guard against misclassification the standard needs to be adjusted by the standard error of the statistic and/or measurement (Livingston & Zieky, 1983; Macready & Dayton, 1980; Poggio, 1984).

   This process is related to #4 above and is problematic insofar as it introduces another need for judgment in the absence of accepted guidelines.

6. Introducing normative information to the standard setting process serves to improve the psychometric characteristics of the method.

   Permitting iteration of the judgment activity, allowing discussion among standard setters, and providing item normative data all have been shown to improve the reliability of the standard-setting method; however, the actual level of the standard most often goes unchanged.

The research reported to date has been thorough in documenting the characteristics of and similarities among methods. Yet there remain numerous issues requiring attention. If fair and equitable standards are to be

realized, further research that can assist practitioners is acutely needed. One need only read the reports from different locales detailing how cut scores were established to be astonished at the diversity of methods being used or the diversity with which the same method is used. Study of these efforts convinces us that although the users were well intended, the research community has failed to be clear, precise, and thorough in presenting the methods and techniques for setting standards. Further, although the topic has been researched for more than a decade, little has been accomplished in terms of exploring alternative methods or giving consideration to the validity of existing methods.

The research that we proposed has the following objectives:

1. to evaluate the validity of the explicit assumptions of commonly used standard-setting methods

2. to develop and test alternate standard-setting methods

3. to examine psychometric characteristics of methods and tests necessary to yield equitable cut scores

4. to prepare a user's handbook on standard setting

Program on Psychometric Theory and Applications

The final direction that resulted from our literature reviews, needs assessments, and discussions centered on the application of psychometric theory of testing problems in schools.

The emphasis in our proposed program of work is on two technical advances of the 1970s and 1980s: item response theory (IRT) and criterionreferenced testing (CRT) (Hambleton & Swaminathan, 1985; Berk, 1984). The first is a methodological advance that can be applied to all types of testing instruments and data. The second is an alternative to norm-referenced testing and provides a basis for looking at students in relation to standards. Both advances are central to what is called modern measurement and have considerable potential for improving testing practices.

We noted first that criterion-referenced tests have more potential for successfully integrating teaching, instruction, and assessment than norm-referenced tests. Also, far less is known about the proper construction and uses of criterion-referenced tests than norm-referenced tests. The latter have been studied since about 1910; the former were only introduced in 1969 (Hambleton, 1982). Second, in view of the potential of item response theory for solving a wide variety of testing problems (Hambleton, 1983), emphasis on this general line of research seemed highly appropriate and consistent with the main direction for testing research today.

Item response theory provided a promising framework for the study of many testing problems, including item bias, test development, individually tailored or adaptive tests, and test score equating. If it can be determined that one or more item response models fit criterion-referenced test data, these models will be useful in helping to address several unique problems that arise in criterion-referenced measurement (CRM): (a) optimal item selection to maximize the decision-making capabilities of short CRTs, (b) the need for cali-

brated banks of test items where item statistics are not student group dependent, and (c) the development of scales for reporting achievement growth. There are also some non-unique problems that arise in criterion-referenced measurement that will require investigation prior to the full implementation of item response models in CRM: (a) assessing model-data fit, (b) estimating model parameters with short test and/or small samples of examines, (c) model selection to solve particular measurement problems, (d) development of new IRT models with more diagnostic capabilities (Embretson, 1985a), and (e) production of developmental scales (e.g. Bergan, 1984).

Criterion-referenced testing technology probably has received even more attention from researchers over the last 15 years than item response theory. Not surprisingly then, there are presently fewer major problems. There appears to be sufficient knowledge available today to build reliable and valid criterion-referenced tests for local schools. Still, important psychometric work remains to be done. Our work focused on problems such as specifying instructional objectives (this needs to be done well, because objectives serve as the "targets" for instruction, and they are central in test development), determining test lengths (classical methods are not applicable), assessing test score and decision validity, and reporting test score information.

Over the last 15 years, much of psychometric research has been directly addressed to real testing problems that arise in schools. We aim to continue that tradition by developing new models and procedures to solve practical measurement problems that were identified by our needs assessments and by our own observations of testing problems being presently faced by school personnel.

The priority projects for the Program on Psychometric theory included:

- Solving Criterion-Reference Measurement Problems with Item Response Models
- Study of Residual Analysis in Test Design
- Advances in Criterion-Reference Test Score Reporting
- Patterns of Item Response

The overall goal of our psychometric research thrust is to enhance the usefulness of criterion-referenced tests to address several problems associated with criterion-referenced tests--choosing test items, selecting test lengths, and adaptively administering tests. These problems are to be addressed using models and procedures within an item response theory framework. In view of the newness of the technology and its infrequent application to criterion-referenced test data, several methodological issues were addressed as part of the research project planning:

1. assessment of model fit
2. estimation of parameters
3. equating test forms
4. assessment of item and test bias

Depending on the results of this work, the development of some new IRT models may be necessary.

The research project is organized around two related components: CRT methodological studies and IRT methodological studies.

This program provides one important theoretical and methodological base for all of the Center's research and development work. In addition to conducting their own projects, it is planned that Drs. Hambleton and Swaminathan will work with the other three programs.

III. The Future of Educational Research and Development in Student Testing, Evaluation, and Standard Setting

The purpose of this paper is to describe, in a non-technical manner, the work needed, over the next five years, to accomplish desirable goals in the area of student testing, evaluation, and standard setting. This paper is the result of a cooperative planning effort undertaken by specialists in student testing, evaluation, and standard setting at Boston College, the Dallas Independent School District, the University of Kansas, the University of Massachusetts, and Western Michigan University. The paper is intended to summarize very briefly where these experts see research and development heading in the near future and how it should impact education. The views presented here are based on extensive literature reviews, interviews, a survey of key informants, and discussions among experts at the cooperating institutions.

Trying to predict the future is problematical at best, however, it is possible to project how present developments in testing will affect education in general at least for the immediate future. This paper addresses futures in the use of tests, in microcomputer-based adaptive testing, in standard setting, and in psychometric theory.

The Use of Tests

The future of research and development on uses of tests should build on past efforts and on the projects described in the WMU Center proposal.[1] There is a substantial body of literature on the uses of tests in educational settings that is not well integrated and which has not been analyzed to establish principles for educational testing that could be used to enhance the utility, feasibility, propriety, and technical quality of testing in any educational setting. The long-term goal of a center on student testing should be to establish a firm knowledge base about educational testing that is then translated into operational guidelines for educators and those who use information about student status and development.

The future of research and development on uses of tests should be directed toward answering the following questions:

1. What organizational forms for testing have been found to be most useful, feasible, proper, and technically sound for each of these different consumer groups: teachers, school building specialists, principals, superintendents, counselors, school boards, state departments of education, legislators, parents, the non-parent community, and policy makers at the national level? Does the success of different organizational forms vary systematically by educational level or within consumer group?

---

[1] These projects include: The use of tests in policy, studies of typical classroom, school, and district uses of testing, examination of testing and standard-setting practices in other countries, a visiting scholars and practitioners project, and school partnerships.

2. What are the information needs of each of the above listed consumer groups and to what extent are these needs being met by existing methods of testing? What needs for information are not being met by existing methods of testing? What alternative approaches to testing are likely to address each unmet need, and how do they work when tried?

The long-term objectives for research and development on the uses of tests should be firmly grounded on past and present research, on reliable and valid data on consumer information needs, and on creative design and prototype testing efforts. Once new approaches have been tested, they should be packaged for dissemination, and demonstration projects should be arranged. Specifically, researchers should:

1. Conduct an exhaustive search for studies of testing used for different purposes, with different consumer audiences, and in different settings. These studies should inform the research staff about what has been tried and about comparative strengths and weaknesses of past practices. These studies should be analyzed for patterns of practice that may form the basis for a set of principles that would guide future work on the use of testing.

2. Conduct an exhaustive search for studies of information needs of different consumer audiences and of the extent to which they have been addressed by available testing methods. These studies should be analyzed for patterns of needs and effective testing methods for reducing these needs. The analysis would form the basis for specifying needs that remain.

3. Conduct input evaluation studies aimed at developing alternative strategies for addressing each unmet need. These studies should be aimed at preparing effective, feasible, proper, and technically sound approaches to meeting the needs of educational consumer groups. These approaches should be systematically tested and the results made available for public discussion and critique.

4. Search for and develop new approaches and support materials for any new approaches to testing that are found to be effective, new means of achieving previously unmet and important needs for identified consumers of information about students.

5. Arrange for informing key actors in education about new approaches with the goal of achieving widespread adoption and use of them in American education.

During the last five years, efforts to reform education, particularly at the state level, have increasingly employed tests and test results in various ways. And this use of tests in the policy sphere is a growing trend. For example, a 50-state survey of reform measures conducted by EDUCATION WEEK (2/6/85) found the following: 29 states require competency tests for students, and 10 other states have such a requirement under consideration; 15 states require an exit test for graduation, 4 additional states have such a measure under consideration; 8 states employ a promotional "gates" test, while 3 others are considering such a mandate; finally, 37 states have some sort of

state assessment program, and 6 additional states have such a program under consideration. In addition to these programs, there are merit pay programs, compensatory funding programs, and teacher testing programs all of which involve the use of test results in reform efforts.

This growing use of tests in the policy sphere cannot but impact on teachers and students, as well as on more traditional testing programs, evaluations of students, and standards of educational excellence. Therefore, it is imperative that the NIE Center, as part of its future work, document the impacts--both positive and negative, the costs and the benefits--associated with these external testing programs. Further, we feel there is a pressing need to develop practical strategies and techniques that state departments of education and local school systems can use to evaluate these programs and to make better use of the information they can provide.

The internal use of tests in schools is equally problematical. There is a convincing body of literature that tells us that millions of dollars are spent on school testing programs that provide information that is frequently unused. The student information needs of local consumers, such as teachers, principals, superintendents, and parents are not well known, and it is difficult to try to compare information generation (through testing) with information needs. Research on the uses of tests in schools must provide the testing industry with knowledge about consumer behavior if the utility of testing for improving student development is to be improved. New reporting systems of perhaps different information than has been traditionally reported potentially can move testing from the periphery to the mainstream of teaching and learning.

Many of the administrative uses of tests have historical counterparts either in this country or abroad. Further, the use of tests as a certification mechanism is widespread in Europe. Therefore, the Center should identify and analyze past uses of tests in policy and in schools both in this country and abroad. This historical analysis would gather evidence on how these programs evolved, how they fit into the structure of the particular educational system, and what their positive and negative effects were on teaching and learning.

Microcomputer-Based Adaptive Testing

Microcomputer-based adaptive testing offers two immediate avenues for considering promising applications. The first comes in the consideration of pragmatic testing concerns. Here such a system has immediate applications in a number of areas that are described below. The second avenue comes from the potential of the system to contribute to research in instruction and the role of testing in refining knowledge about instruction. Both will be able to be pursued upon the successful completion of research and development on microcomputer-based adaptive testing, and both offer high potential for realizing great dividends.

A system to be developed by the Center would provide microcomputer-based software which allows for the selection of items from an indexed bank for adaptive testing. The system would have links to an instructional hierarchy which would help teachers with the selection of instructional approaches with individual students or classes through the testing output. These two features of the system provide its future potential.

A severe limitation to repeated testing in classrooms is the high cost, both in terms of time and money. Furthermore, to be useful, repeated testing requires rapid feedback to the teacher and student. The practical advantage of the proposed system is that the cost of the system is minimal when spread over the large number of students that can access it and the increased testing throughput that becomes possible. Repeated testing becomes feasible under either criterion. When considered in light of testing time, the testing can be accomplished rapidly with the use of adaptive item selection and the relatively fast response times of a state-of-the-art microcomputer. Furthermore, the feedback for the student and teacher comes in minutes and seconds instead of weeks and days. The feedback can be instantly tailored for any size group of students or the results for a large group can be used to form more effective subgroups for instructional purposes. Finally, the reduced testing time makes increased use of the system possible without an undue detraction from instructional time.

Another important practical advantage of the system is the ultimate potential for its easy adaptation for use with new technologies such as video disc or expanded-capability terminals. A video disc-based item delivery system could be efficiently and effectively merged with the system proposed, as could more versatile graphics terminals. Systems such as these offer great potential for employing new item form and testing situations.

From the research perspective, the system offers many possibilities. A proposed initial base for the system is mathematics. Mathematics was chosen because of the large number of existing curricula and satisfactory item pools. The instructional linkage with these existing materials can be straightforward and immediately usable in the classroom.

The research potential with this existing system will be great. As one example, research in mathematics instruction and testing has shown that the options a student selects in answering a problem give insight into his or her conceptual approach to the task. Knowledge of the responses to the items can be easily tracked for students and groups of students allowing teachers and mathematicians to analyze the relationship between different instructional approaches and the degree to which tasks are mastered.

Greater potential comes in utilizing the testing capabilities of microcomputer-based adaptive testing in deciphering the hierarchical structures of more complex skills such as reading. The field of reading research has been plagued by the inability to determine effective hierarchical structures for teaching reading. The literature is rife with studies which directly contradict each other. The advantages offered by the proposed system in determining effective structures and the concomitant teaching methodologies are immense. The possibility of immediate testing and of having an instructional analysis system which can instantly match students and their known characteristics with their performance under different instructional systems opens tremendous opportunities to advance reading research.

In short, the long term potential of microcomputer-based adaptive testing is unlimited in terms of both immediate practical application and future advances in research utilizing testing in the analysis of effective instruction.

Standards Setting

The common thread that runs through the future of research and development in standard-setting is an understanding of how standards evolve and take hold in society, and an examination of methods that yield equitable test performance standards. Further, the research programs to which we are committed are, for the most part, tied to providing information for school-based practitioners. We are committed to a mission that views testing and evaluation as important elements in the teaching and learning process, but recognizes testing and evaluation as tools in this process.

Through our intended research agenda we plan to undertake programs that enhance the fair and equitable use of tests at the local level. Thrusts that we envision for research and development on standard setting over the next five years are:

1. Standards - Establishing Passing Scores

   Based on our work and others', we do not believe "truth" will be discovered in this arena. That is, a discovery of the method to yield the passing score is quixotic. Performance distributions are continuous and unimodal. The goal of a standard setting method is to provide an objective, reliable, and valid procedure for securing the values of judges in such a way that the range of possible passing scores is narrowed. From this restricted boundary, discussion leading to consensus can occur. Recognizing fallibility as a premise or precondition, inquiries should be conducted in such areas as:

   a. validity characteristics of a host of available methods;

   b. defining other methods - we are particularly interested in combining concerns tied to curricular and instructional validity to setting the cut score, and exploring scaling methods (Thurstone, Guttman);

   c. alternate standards configurations, e.g., total score-versus objective passing plus total score, test dimensionality and test length and the accuracy of classification, student population characteristics (e.g. handicapped) and the usefulness of information;

   d. inquiry into teacher-versus-test classification, examining sources of discrepancy, consequences, test use, test credibility, test cognitive dissonance, etc.;

   e. possibility of local standards versus imposed standards (SEA) tied to test use, impact, and policy;

   f. issues of equity linked to students close to the passing score, potentially broadening discrimination into 3 categories (at least), e.g., failers, remedials, and passers; and,

   g. isolation and documentation of factors that effect setting test performance tandards.

2. Standards - Impact of Testing and Evaluation on Societal Expectations and Implementation

The research agenda in this area would be less likely geared to experimental/quasi-experimental studies. Historical as well as naturalistic forms of inquiry are likely to predominate. Of interest would be:

a. inquiry on the political aspects and realities of evaluation users;

b. questions of believability, importance, and use of evaluation and testing products by key decision makers; and,

c. characteristics of processes, products, and deliverables that effect standards formation, selection, adoption, and incorporation.

3. Curriculum-Test Match

The issues here are quickly becoming apparent. Seminal studies are needed. Findings need to be communicated to users. Further, there is a need to create, then establish, designs that evaluate fit/match. Studies being considered include:

a. impact on standards and adequate adjustments;

b. extent of content taught but not tested;

c. evaluating school effectiveness in consideration of conditions imposed by curriculum-test match.

Psychometric Theory

Presently, we can see at least three prominent directions for the research:

1. criterion-referenced testing methodology,
2. computers and testing,
3. cognitive theory and psychometrics.

A very brief description of each research direction follows:

1. Criterion-Referenced Testing Methodology

Our research would have three principal objectives:

a. to develop some new methods for determining criterion-referenced test lengths. Of main interest would be a method that involves practitioners using the computer to simulate realistic test results so as to empirically investigate the effects of a number of factors (i.e., cut-off score, score distribution) prior to choosing a test length.

b. to investigate the use of optimal item selection on the decision-making effectiveness of tests, and the consequences of optimal item selection on content validity and on the motivational and psychological influence of statistically similar test items;

c. to design and field test several new methods for assessing criterion-referenced test score validity. Of special interest would be decision-theoretic approaches involving new functions and consideration of problems in choosing samples. The results from this research would include some specific guidelines for validating criterion-referenced tests which are consistent with the 1985 APA/ AERA/NCME Test Standards.

The results from the above three research studies, and related studies, can be of considerable value to school and state testing personnel who have the task of building technically sound and defensible criterion-referenced tests. The research studies should be carried out in a coordinated way with participating school districts and state departments, so that the final results will be understandable and useful to these important users of criterion-referenced tests.

2. Computers and Testing

We predict that in the coming years, the present computer revolution in this country will greatly influence the ways in which educational and psychological tests are developed, administered, scored, and interpreted. To date, the impact of computers on testing practices has been limited to the uses of computers in item banking and adaptive testing with multiple-choice test questions.

The principal goal of our research program would be to enhance the validity of exam scores and associated decisions by effectively using main-frame and micro-computers. Specifically, our research would center on the use of free-response questions, video-disc technology, sequencing problems, new scoring formats, and other testing innovations that can be addressed with computer technology. To date, almost no research along the general lines described above has been conducted.

3. Cognitive Theory and Psychometrics

One of the frontier areas of psychometrics is the merger between modern cognitive theory and psychometrics. Within the framework of cognitive theory, new important variables that influence learning and retention are being identified. These variables need to be fully defined and measured, and validity studies must be carried out. Also, new psychometric models, such as some of the new multi-cognitive components models being developed and studied by Fischer , Embretson, and others, must be further developed. Our principal goal would be to build on the existing research by developing new psychometric models based on some of our other item response model research to facilitate development and analysis of theory-based tests, and by demonstrating the relevance of these new psychometric models and tests in educational settings.

In summary, research and development at a National Center on student testing, evaluation, and standard setting should be building a knowledge base in important areas where little is currently known. It should translate this new knowledge into workable models that can be used in classrooms, schools and school districts, and for state and national level reform. It should pioneer new advances which may not have direct impact on educational practices in the short term, but would expand the research and development landscape for future generations. And, school districts should be directly involved in designing, conducting, and applying the needed research and development.

BIBLIOGRAPHY FOR PLANNING GRANT

Adler, M. J. (1982). **The Paideia proposal: An education manifesto.** New York: Macmillan.

Airasian, P. W., Kellaghan, T., Madaus, G. F., & Pedulla, J. J. (1977). Proportion and direction of teacher rating changes of pupils' progress attributable to standardized test information. **Journal of Educational Psychology,** 69, 702-709.

Airasian, P. W., & Madaus, G. F. (1983). Linking testing to instruction: Policy issues. **Journal of Educational Measurement,** 20, 103-118.

Alexander, C. (1985). **Helping classroom teachers use tests and testing results.** Unpublished manuscript, Dallas Independent School District, Department of Research, Evaluation, and Information Systems.

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

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for examination standards. **Educational and Psychological Measurement,** 36, 45-50.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), **Educational measurement.** Washington, DC: American Council on Education.

Archer, P. (1979). **A comparison of teacher judgments of pupils and the results of standardized tests.** Unpublished doctoral dissertation. University of College Cork, Ireland.

Arns, R. G., & Urban, P. A. (1984). Strategic choices for data communications systens, **Cause/Effect,** 7 (5), 6-12.

Austin, G. R. (1981). Exemplary schools and their identification. In D. Carlson (Ed.), **New directions for testing and measurement.** San Francisco: Jossey-Bass.

Austin, G. R., Chafin, A. E., Hambleton, R. K., Stufflebeam, D. L., Garber, H., & Gordon, C. H. (1985, April). **Evaluation of a statewide CBT program from different perspectives.** Symposium conducted at the meeting of the American Educational Research Association, Chicago.

Bartell, N. R., Grill, J. J., & Bryen, D. N. (1973). Language characteristics of black children: Implications for assessment. **Journal of School Psychology,** 11, 351-364.

Bates, A. W. (1984, September). **The implications for teaching and learning of new informatics developments** (I.E.T. Papers on Broadcasting No. 233). Paper presented at the Annual Conference of Higher Education International, York, England. (ERIC Document Reproduction Service No. ED 253 208)

Beck, M. D. & Stetz, F. P. (1979, April). **Teachers' opinions of standardized test use and usefulness.** Paper presented at the meeting of the American Educational Research Association, San Francisco.

Beggs, D. L., Mayer, G. R., & Lewis, E. L. (1972). The effects of various techniques of interpreting test results on teacher perception and pupil achievement. **Measurement and Evaluation in Guidance,** 5, 290-297.

Bergan, J. (1984). **Head start measurement battery.** Final Report. Submitted to the Department of Health and Human Services.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. **Journal of Experimental Education,** 45, 4-9.

Berk, R. A. (1984). **A guide to criterion-referenced test construction.** Baltimore: Johns Hopkins University Press.

Berk, R. A. (1985, April). **A consumers guide to setting performance standards on criterion reference tests.** Paper presented at the meeting of the National Council on Measurement in Education, Chicago.

Bettinghaus, E. P., & Miller, G. R. (1973). **A dissemination system for state accountability programs, Part II: The relationship of contemporary communication theory to accountability dissemination theories.** Denver, CO: Cooperative Accountability Project.

Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. In R. W. Tyler (Ed.) **Educational evaluation: New roles, new means. The sixty-eighth yearbook of the National Society for the Study of Education, Part II.** Chicago: NSSE.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). **Taxonomy of educational objectives: The classification of educational goals. Handbook 1. Cognitive domain.** New York: David McKay.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). **Handbook on formative and summative evaluation of student learning.** New York: McGraw-Hill.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. **Psychometrika,** 46, 443-459.

Bosma, B. (1973). The NEA testing moratorium. **Journal of School Psychology,** 11, 304-306.

Boyer, E. L. (1983). **High school: A report on secondary education in America.** New York: Harper & Row.

Brickell, H. M. (1976). The influence of external political factors on the role and methodology of evaluation. **Educational Comment,** 5 (2), 1-6.

Brim, O. G., Jr., Glass, D. C., Neulinger, J., Firestone, I. R., & Lerner, S. C. (1969). **American beliefs and attitudes about intelligence.** New York: Russell Sage Foundation.

Broadfoot, P. (Ed.). (1984). **Selection, certification, & control: Social issues in educational assessment.** Barcombe Lewes, Sussex, England: The Falmer Press.

Brookover, W. B. (1959). A social psychological conception of classroom learning. **School and Society,** 87, 84-87.

Brumm, L. (1983, December). **Delivering technical education in Wisconsin in the information age.** Paper presented at the American Vocational Association Convention, Anaheim, CA. (ERIC Document Reproduction Service No. ED 237 787)

Bunda, M. A. (1985). **Influence of training in measurement skills in higher education.** Unpublished manuscript, Western Michigan University, Evaluation Center, Kalamazoo.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. **Psychological Bulletin,** 56, 81-105.

Chall, J. (1977). **An analysis of textbooks in relation to declining SAT scores.** New York: College Entrance Examination Board.

Clark, K. B. (1963). Educational stimulation of racially disadvantaged children. In A. H. Passow (Ed.), **Education in depressed areas.** New York: Bureau of Publications, Columbia University.

College Entrance Examination Board. (1977). **On further examination: Report of the advisory panel on the scholastic aptitude score decline.** New York: Author.

Committee to Develop Standards for Educational and Psychological Testing of The American Educational Research Association, The American Psychological Association, and The National Council on Measurement in Education. (1985). **Standards for educational and psychological testing.** Washington, D. C.: American Psychological Association.

Connell, C. (1978, November 12). The going gets tough for educational testers. **The Boston Globe,** C16.

Cooley, W. (1985). **Computer assisted professional: A proposal to develop an information system to assist school professionals in planning and implementing school improvement.** Unpublished manuscript, University of Pittsburgh, Learning Research and Development Center.

Cronbach, L. J. (1963). Course improvement through evaluation. **Teachers College Record,** 64, 672-83.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. **American Psychologist,** 30, 1-14.

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

De Bevoise, W. (Ed.). (1983). **Collaboration wears a layered look.** Eugene: University of Oregon, Center for educational Policy and Management. (ERIC Document Reproduction Service No. ED 238 128)

de Gruijter, D. N. M., & Hambleton, R. K. (1983). Using logistic test models in criterion-referenced test item selection. In R. K. Hambleton (Ed.), **Applications of item response theory.** Vancouver, BC: Educational Research Institute of British Columbia.

deRivera, M. (1974). Testitis: A technical affliction. **Childhood Education,** 50, 217-221.

Dreeben, R. (1969). **On what is learned in schools.** Reading, MA: Addison-Wesley.

Dwyer, M. M. (1984). **Indiana partners in education handbook.** Indianapolis: Hoosiers for Economic Development Committee.

Ebel, R. L. (1979). **Essentials of educational measurement** (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Education Commission of the States, Task Force on Education for Economic Growth. (1983). **Action for excellence.** Denver: ECS.

Embretson, S. (Ed.). (1985, a). **Test design.** New York: Academic Press.

Embretson, S. (1985, b). **Test design: Cognative models of item response.** Unpublished manuscript, University of Kansas, Lawrence.

Florida Phi Delta Kappa Consortium Planning Task Force. (1985). **Evaluating the impact of educational reforms in Florida.** Gainesville, FL: Author.

Flowers, C. E. (1966). **Effects of an arbitrary accelerated group placement on the tested academic achievement of educationally disadvantaged students.** Unpublished doctoral dissertation, Teacher's College, Columbia University, New York.

Fox, M. R., Faver, C. A. (1984). Independence and cooperation in research. The motivations and costs of collaboration. **Journal of Higher Education.** 55, 347-359.

Frisbie, R. D., & Thompson, T. L. (1985). **The program on technology in student testing, evaluation, and standards.** Unpublished manuscript, Western Michigan University, Evaluation Center, Kalamazoo.

Fullan, M., & Pomfret, A. (1977). Research on curriculum and instruction implementation. **Review of Educational Research,** 47, 335-397.

Gay, G. & Abrahams, R. D. (1973). Does the pot melt, boil, or brew? Black children and white assessment procedures. **Journal of School Psychology,** 11, 330-340.

Glaser, B. (1978). **Theoretical sensitivity.** Mill Valley, CA: Sociology Press.

Glaser, B., & Strauss, A. (1967). **The discovery of grounded theory.** Chicago, IL: Aldine.

Glasnapp, D. (1985). **Criterion-referenced testing: Research directions.** Unpublished manuscript, University of Kansas, Center for Educational Testing and Evaluation, Lawrence.

Glass, G. V. (1978). Standards and criteria. **Journal of Educational Measurement,** 15, 237-261.

Gonzalez, M. L. (1985). **Factors affecting the utility of standardized tests.** Unpublished manuscript, Dallas Independent School District, Department of Research, Evaluation, and Information Systems.

Goodlad, J. (1983). **A Place Called School.** New York: Harper & Row.

Goslin, D. A. (1963). **The search for ability: Standardized testing in social perspective.** New York: Russell Sage Foundation.

Goslin, D. A. (1967). **Teachers and testing.** New York: Russell Sage Foundation.

Goslin, D.A. & Glass, D. C. (1967). The social effects of standardized testing in American elementary and secondary schools. **Sociology of Education,** 40, 115-131.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive testing. **Journal of Educational Measurement,** 21, 347-360.

Green, R. L. (1975). Tips on educational testing: What teachers and parents should know. **Phi Delta Kappan,** 57, 89-93.

Guba, E. G. (1982, April). **The search for truth: Naturalistic inquiry as an option.** Paper presented at a meeting of the American Educational Research Association, Chicago.

Guba, E. G., & Lincoln. Y. S. (in press). Do inquiry paradigms imply inquiry methodologies? In David Fetterman (Ed.), **(Title Undetermined),** Beverly Hills, CA: Sage.

Hambleton, R. K. (1982). Advances in criterion-referenced testing technology. In C. Reynolds & T. Gutkin (Eds.), **Handbook of School Psychology.** New York: Wiley.

Hambleton, R. K. (Ed.). (1983). **Applications of item response theory.** Vancouver, B.C.: Educational Research Institute of British Columbia.

Hambleton, R. K. (1984). Determining suitable test lengths. In R. Berk (Ed.), **Criterion-referenced measurement: State of the art.** Baltimore: The Johns Hopkins University Press.

Hambleton, R. K., Anderson, G. E., & Murray, L. (1983). Applying microcomputers to classroom testing practices. In W. Hathaway (Ed.), **New directions for testing and measurement: Testing in the schools.** San Francisco: Jossey-Bass.

Hambleton, R. K., & de Gruijter, D. N. M. (1983). Application of item response models to criterion-referenced test item selection. **Journal of Educational Measurement,** 20, 355-367.

Hambleton, R. K. & Murray, L. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), **Applications of item response theory** Vancouver, BC: Educational Institute of British Columbia.

Hambleton, R. K., & Swaminathan, H. (1985). **Item response theory: Principles and applications.** Boston: Kluwer-Nijhoff.

Haney, W. (1984). Testing reasoning and reasoning about testing. **Review of Educational Research,** 54, (forthcoming).

Hein, G. E. (1975). Standardized testing: Reform is not enough. In M. Cohen (Ed.), **Testing and evaluation's new views.** Washington, D. C.: Association for Childhood Education International.

Heiry, T. J. (1985). **Position paper on assessment for special education students.** Unpublished manuscript, Dallas Independent School District, Department of Research, Evaluation, and Information Systems.

Herndon, T. (1975). **Standardized tests: Are they worth it?** Paper presented to the Commonwealth Club, San Francisco.

Hoffman, B. (1962). **The tyranny of testing.** New York: Crowell-Collier Press.

Holmen, M. G., & Doctor, R. F. (1972). **Educational and psychological testing: A study of the industry and its practices.** New York: Sage.

Holt, J. (1968). **On testing.** Cambridge, MA: Pinck Leodas Association.

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

House, E., Rivers, W., & Stufflebeam, D. (1974). An assessment of the Michigan accountability system. **Phi Delta Kappan,** 55, 663-669.

Houts, P. L. (1975). A conversation with Banesh Hoffman. **National Elementary Principal,** 54 (6), 2-3.

Houts, P. L. (1977). Introduction: Standardized testing in America. In P. L. Houts (Ed.), **The myth of measurability.** New York: Hart Publishing.

Huling, L. L., Richardson, J. A., & Hord, S. M. (1983). Three projects show how university/school partnerships can improve effectiveness. NASSP **Bulletin**, 67 (465), 39-44.

Husserl, E. (1969). **Formal and transcendental logic** (Dorian Cairns, Trans.). The Hague: Martinus Nijhoff.

Jackson, P. (1968). **Life in the classroom.** New York: Holt, Rinehart, & Winston.

Jaeger, R. M. (1976). Measurement consequences of selected standard-setting models. **Florida Journal of Educational Research**, 18, 22-27.

Jaeger, R.M. (1978, Spring). **A proposal for setting a standard on the North Carolina High School Competency Test.** Paper presented at the meeting of the North Carolina Association for Research in Education, Chapel Hill.

Jaeger, R. M., & Tittle, C. K. (Eds.), (1980). **Minimum competency achievement testing: Motives, models, measures, and consequences.** Berkeley, CA: McCutchan.

Jensen, A. R. (1980). **Bias in mental testing.** New York: The Free Press.

Joint Committee on Standards for Educational Evaluation. (1981). **Standards for evaluations of educational programs, projects, and materials.** New York: McGraw-Hill.

Jones, A. H., & Barnes, C. P. (1984). The California consortium: A case study on seeking change in teacher education. **Journal of Teacher Education**, 35 (6), 5-10.

Kellaghan, T., Madaus, G. F., & Airasian, P. W. (1982). **The effects of standardized testing.** Boston: Kluwer-Nijhoff Publishing.

Kennedy, M., Apling, R., & Neumann, W. (1980). **The role of evaluation and test information in public schools.** Cambridge, MA: The Huron Institute.

King, J. A., & Pechman, E. M. (1982). **The process of evaluation use in local school settings** (Final report of NIE grant 81-0900). New Orleans: New Orleans Public Schools.

Kirkland, M. C. (1971). The effects of tests on students and schools. **Review of Educational Research**, 41, 303-350.

Kottler, S. L. (1980). A comparison of approaches for setting proficiency standards. **Journal of Educational Measurement**, 17, 167-178.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). **Taxonomy of educational objectives: The classification of educational goals. Handbook 2: Affective domain.** New York: David McKay.

Kreitzberg, C. B., & Jones, D. H. (1980). **An empirical study of the broad-range tailored test of verbal ability** (RR-80-5). Princeton, NJ: Educational Testing Service.

Lazarus, M. (1975). Coming to terms with testing. **National Elementary Principal**, 54 (6), 24-29.

Leiter, K. C. W. (1976). Teachers' use of background knowledge to interpret test scores. **Sociology of Education**, 49, 59-65.

LeMahieu, P. G. (1984). The effects on achievement and instructional content of a program of student monitoring through frequent testing. **Educational Evaluation and Policy Analysis**, 6, 175-187.

Lewandowski, A. R. (1984). Implementing an information center in a complex university environment. **Cause/Effect**, 7 (1), 6-9.

Lewis, B. (1977, November). Testing: A parent's point of view. In R. M. Bossone & M. Weiner (Eds.), **Proceedings from the National Conference on Testing: Major issues.** New York: Center for Advanced Study in Evaluation, City University of New York

Lincoln, Y. S., & Guba, E. G. (1985). **Naturalistic inquiry.** Beverly Hills, CA: Sage.

Lincoln, Y. S., & Guba, E. G. (in press). But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. Chapter 3 in David Williams (Ed.), **Sourcebook on program evaluation: New directions for program evaluation.** San Francisco, CA: Jossey-Bass.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). An investigation of item bias in a test of reading comprehension. **Applied Psychological Measurement**, 5, 159-173.

Linn, R. L., Madaus, G. F., & Pedulla, J. J. (1982). Minimum competency testing: Cautions on the state of the art. **American Journal of Education**, 91, 1-35.

Livingston, S. A., & Ziecky, M. J. (1983). **Passing scores: Manual for setting standards of performance in educational and occupational tests.** Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). **Application of item response theory to practical testing problems.** Hillsdale, NJ: Lawrence Erlbaum Associates.

Lortie, D. (1975). **Schoolteacher.** Chicago: University of Chicago Press.

Lutjeharms, J. (1983). **State dissemination grants program, educational research and development: February, 1978 through June, 1983** (Final Report, NIE-G-75-0021). Lincoln: Nebraska State Department of Education. (ERIC Document Reproduction Service No. ED 251 939)

Lyon, C. (1978, July). **What do we know about teaching and learning in urban schools? Vol. 9 New perspectives on school district research and evaluation.** Paper presented at the National Conference on Urban Education, St. Louis, MO.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. **Journal of Educational Statistics,** 2, 99-120.

Macready, G. B., & Dayton, C. M. (1980). The nature and use of state mastery models. **Applied Psychological Measurement,** 4, 493-516.

Madaus, G. F. (1981). NIE Clarification Hearing: The negative team's case. **Phi Delta Kappan,** 63 (2), 92-94.

Madaus, G. F. (Ed.). (1982). **The courts, validity, and minimum competency testing.** Boston: Kluwer-Nijhoff.

Madaus, G. F. (1985, a). **Public policy and the testing profession -- You've never had it so good?** Presidential address at the annual meeting of the National Council on Measurement in Education, Chicago.

Madaus, G. F. (1985, b). **Review of the effects of standardized testing.** Unpublished manuscript, Boston College, Center for the Study of Student Testing, Evaluation, and Educational Policy.

Madaus, G. F. (1985, c). **Standards for educational and psychological testing.** Unpublished manuscript, Boston College, Center for the Study of Student Testing, Evaluation, and Educational Policy.

Madaus, G. F. (1985, d). Test scores as administrative mechanisms in educational policy. **Phi Delta Kappan,** 66, 611-617.

Madaus, G. F. (1985, e). **Use of tests in policy.** Unpublished manuscript, Boston College, Center for the Study of Student Testing, Evaluation, and Educational Policy.

Madaus, G. F., Airasian, P. W., & Kellaghan, T. (1980). **School Effectiveness: A reassessment of the evidence.** New York: McGraw-Hill.

Madaus, G. F. & Greaney, V. (1985). The Irish experience in competency testing: Implications for American education. **American Journal of Education,** 93, 268-294.

Madaus, G. F., Scriven, M., & Stufflebeam, D. L. (1983). **Evaluation models: Viewpoints on educational and human services evaluation.** Boston: Kluwer-Nijhoff.

Madaus, G. F., & Stufflebeam, D. L. (1984). A review of efforts to assure the quality of education through program evaluation and accountability. **American Behavioral Scientist,** 27, 649-672.

McKenna, B. H. (1975). A tale of testing in two cities. **National Elementary Principal,** 54 (6), 40-45.

Mead, R. J. (1975). **Analysis of Fit to the Rasch Model.** Unpublished doctoral dissertation, The University of Chicago.

Meier, D. (1973). **Reading failure and the tests.** Occasional paper of the Workshop Center for Open Education, New York.

Mendro, R. (1985). **Issues in testing in bilingual education.** Unpublished manuscript, Dallas Independent School District, Department of Research, Evaluation, and Information Services.

Mercer, J. R. (1973). **Labeling the mentally retarded.** Berkeley, CA: University of California Press.

Meskauskas, J. A. (1976). Evaluation models for criterion referenced testing: Views regarding mastery and standard setting. **Review of Educational Research,** 45, 133-158.

Metallinos, N. (1984, May). **Approaches to human communication training: The sociological focus.** Paper presented at the Delphi Symposium on Developing Human Resources in Communication through University Training, Delphi, Greece. (ERIC Document Reproduction Service No. ED 248 566)

Metfessel, N. S., & Michael, W. B. (1967). A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs. **Educational and Psychological Measurement,** 27, 931-943.

Meyen, Edward. (1985). **Implications of special needs populations and assessment practices.** Unpublished manuscript, University of Kansas, Center for Educational Testing and Evaluation, Lawrence.

Miller, M. D. (1985). **Patterns of item response.** Unpublished manuscript, University of Kansas, Center for Educational Testing and Evaluation, Lawrence.

Millman, J. (1973). Passing scores and test lengths for domain referenced measures. **Review of Educational Research,** 43, 205-216.

Millman, J., & Arter, J. A. (1984). Issues in item banking. **Journal of Educational Measurement,** 21, 315-330.

Mills, C. N., & Simon, R. (1981). **A method of determining the length of criterion-referenced tests using reliability and validity indices** (Laboratory of Psychometric and Evaluative Research Report #110). Amherst: University of Massachusetts, School of Education.

Naisbitt, J. (1982). **Megatrends: Ten new directions transforming our lives.** New York: Warner Books.

National Coalition of Advocates for Students. (1985). **Barriers to excellence: Our children at risk.** Boston: Author.

National Commission on Excellence in Education. (1983). **A nation at risk: The imperative for educational reform.** Washington, D.C.: U.S. Government Printing Office.

National Research Council. (1977). **The state of school science: a review of the teaching of mathematics, science, and social studies in American schools, and recommendations for improvements.** Washington, DC: National Academy of Sciences.

Nedelsky, L. (1954). Absolute grading standards for objective tests. **Educational and Psychological Measurement,** 14, 3-19.

New Jersey Education Association. (1979). **Procedures for the evaluation of the performance of each public school district and school.** New Jersey: Author.

Nibley, A. M. (1979, January 7). The evils of testing. **The Boston Globe,** A11-A12.

Oliver, D. (1983). Deciphering electronic mail: Connecting and interconnecting services. **Library Hi Tech,** 1 (2), 33-48.

Olson, G. H. (1985). **Computer-generated, personalized testing.** Unpublished manuscript, Dallas Independent School District, Department of Research, Evaluation, and Information Services.

Owen, D. (1985). **None of the above: Behind the myth of scholastic aptitude.** Boston: Houghton Mifflin.

Owen, T. (1973). Educational evaluation by adversary proceeding. In E. House (Ed.), **School evaluation: The politics and process.** Berkeley: McCutchan.

Parsons, T. (1959). The school class as a social system: Some of its functions in American society. **Harvard Educational Review,** 29, 297-318.

Perrone, V. (1977). **Alternatives to standardized testing.** Bloomington, IN: Phi Delta Kappa.

Picus, L., & Holznagel, D. (1983, June). **Electronic communication networks for education: Policy implications for SEA's** (Discussion draft). Portland, OR: Nowthwest Regional Educational Laboratory and Northwest Center for State Educational Policy Studies. (ERIC Document Reproduction Service No. ED 252 981)

Pidgeon, D. A. (1970). **Expectation and pupil performance.** Slough Bucks, England: NFER Publishing.

Poggio, J. P. (1984). **Practical considerations when setting test standards: A look at the process used in Kansas.** Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Poggio, J. (1985). **Setting standards.** Unpublished manuscript, University of Kansas, Center for Educational Testing and Evaluation, Lawrence.

Poggio, J. P., & Glasnapp, D. R. (1980). **Report of research findings: The Kansas competency testing program - 1980.** Topeka, KS: Kansas State Department of Education.

Poggio, J. P., Glasnapp, D. R., & Eros, D. S. (1981). **An empirical investigation of the Angoff, Ebel, and Nedelsky standard setting methods.** Paper presented at the meeting of the American Educational Research Association, Los Angeles.

Poggio, J. P., Glasnapp, D. R., & Eros, D. S. (1982). **An evaluation of contrasting groups methods for setting test standards.** Paper presented at the annual meeting of the American Educational Research Association, New York.

Poggio, J. P., Glasnapp, D. R., & Eros, D. S. (1983). **An analysis of the validity of judgmental methods used to set test standards.** Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Popham, W. J. (1978). As always, provocative. **Journal of Educational Measurement,** 15, 297-300.

Popham, W. J. (1981). The case for minimum competency testing. **Phi Delta Kappan,** 63 (2), 89-91.

Popham, W. J., Kruse, K. L., Rankin, S. C., Sandifer, P. D., & Williams, P. L. (1985). Measurement driven instruction: It's on the road. **Phi Delta Kappan,** 66, 628-634.

Quinto, F. (1977, November). Why standardized tests fail the accountability test. In R. M. Bossone & M. Weiner (Eds.), **Proceedings from the National Conference on Testing: Major Issues.** New York: Center for Advanced Study in Education, Graduate School and University Center of the City University of New York.

Radwin, E. (1981). **A case study of New York City: Citywide reading testing program.** Cambridge, MA: The Huron Institute.

Raudenbush, S. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction. **Journal of Educational Psychology,** 76, 85-97.

Resnick, D. P. (1982). History of educational testing. In A. Wigdon & W. Garner (Eds.), **Ability testing** (pp. 173-194). Washington, D. C.: National Academy Press.

Resnick, D. P., & Resnick, L. B. (1985). Standards, curriculum, and performance: A historical and comparative perspective. **Educational Researcher,** 14 (4), 5-20.

Resnick, L. B. (1977, September-October). Matching tests with goals. **Social Policy,** 4-10.

Rice, J. M. (1897). The futility of the spelling grind. **The Forum,** 23, 163-172.

Richman, C. L., Brown, K. P., & Clark, M. (Undated). **Personality changes as a function of minimum competency test success/failure.** NIMH Grant PHS 1R01 Mh36491. (Request for reprints may be made through C. L. Richman, Wake Forest University, Winston-Salem, NC.)

Rist, R. C. (1970). Student social class and teachers' expectations: The self-fulling prophecy in ghetto education. **Harvard Educational Review,** 40, 411-451.

Rist, R. C. (1977). On understanding the processes of schooling: The contribution of labeling theory. In J. Karabel & A. H. Halsey (Eds.), **Power and ideology in education.** New York: Oxford University Press.

Root, D. (1985). **Case study handbook: Partners in education program - Indianapolis, Indiana.** Kalamazoo, MI: Western Michigan University, The Evaluation Center.

Rosenthal, R. & Jacobson, L. (1968). **Pygmalion in the classroom.** New York: Holt, Rinehart, & Winston.

Rossi, P. H. (Ed.). (1982). Standards for evaluation practice. **New directions for program evaluation,** No. 15. San Francisco: Jossey-Bass.

Rozanski, M., & Kelleher, A. (1983). International education consortia: A case study. **Educational Research Quarterly,** 8, 100-107.

Rudman, H. C., Kelley, J. L., Wenous, D. S., Mehrens, W. A., Clark, C. M., & Porter, A. C. (1980). **Integrating assessment with instruction: A review (1922-1980)** (Research Series No. 75). East Lansing, MI: Michigan State University, College of Education, Institute for Research on Teaching.

Ryan, C. (1979, January). **The testing maze.** A national PTA white paper.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: What's really happening? **Phi Delta Kappan,** 62, 631-633.

Samuda, R. J. (1975). **Psychological testing of American minorities: Issues and consequences.** New York: Dodd, Mead & Co.

Samuda, R. J. (1977, November). Critical concerns in the testing of minorities: Time for new initiatives. In R. M. Bossone & M. Weiner (Eds.), **Proceedings from the National Conference on testing: Major issues.** New York: Center for Advanced Study in Edcuation, City University of New York.

Sanders, J. R. (1985). **The natural use of student testing and evaluation in schools by classroom teachers and principals, and how that use may be enhanced to improve teaching and learning at the local level.** Unpublished manuscript, Western Michigan University, Evaluation Center, Kalamazoo.

Sanders, J. R. & Goodwin, W. L. (1971). **Exploring the effects of selected variables in teacher expectation of pupil success.** Unpublished manuscript, Bucknell University, Lewisburg, PA. (ERIC Document Reproduction Service No. EJ 080 591)

Scanlon, R. L. (1973). The perceptual press of classroom constraints. **Irish Journal of Education,** 7, 29-39.

Schambier, R. F. (1983, November). **Staff development: The carrot or the stick?** Paper presented at the annual meeting of the American Association for Adult and Continuing Education, Philadelphia. (ERIC Document Reproduction Service No. ED 237 658)

Schrader, W. (Ed.). (1979). **New directions in testing and measurement** (No. 1). San Francisco: Jossey-Bass.

Schwartz, J. L. (1975). Math tests. **National Elementary Principal,** 54 (6), 67-71.

Scriven, M. (1967). The methodology of evaluation. In **Perspectives on curriculum evaluation** (AERA Monograph Series on Curriculum Evaluation, No. 1). Chicago: Rand McNally.

Scriven, M. (1973). Goal-free evaluation. In E. House (Ed.), **School evaluation: The politics and process.** Berkeley, CA: McCutchan.

Serebraikoff, V. & Langer, S. (1977). Are IQ tests immoral? Have they been debunked? **Your child's IQ.** New York: David McKay.

Shepard, L. A. (1983). The role of measurement in educational policy: Lessons from the identification of learning disabilities. **Educational Measurement: Issues and Practice,** 2 (3), 4-8.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test item bias with both internal and external ability criteria. **Journal of Educational Statistics,** 6, 317-375.

Simon, B. (1971). **Intelligence, psychology, and education. A Marxist critique.** London: Lawrence and Wishart.

Sizer, T. R. (1984). **Horace's compromise: The dilemma of the American high school today.** Boston, MA: Houghton-Mifflin.

Skakun, E. N., & Kling, S. (1980). Comparability of methods for setting standards. **Journal of Educational Measurement,** 17, 229-235.

Smith, N. L. (1981). **New techniques for evaluation** (New Prespectives in Evaluation Series No. 2). Beverly Hills: Sage.

Sproull, L. & Zubrow, D. (1981). Standardized testing from the administrator's perspective. **Phi Delta Kappan,** 62, 628-630.

Stake, R. E. (1975). **Program evaluation, particularly responsive evaluation.** (Occasional Paper Series No. 5). Kalamazoo, MI: Western Michigan University, Evaluation Center.

Stedman, L. C. & Smith, M. S. (1983). Recent reform proposals for American education. **Contemporary Education Review,** 2 (2), 85-104.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. **Applied Psychological Measurement**, 7, 210-210.

Strenio, A. J., Jr. (1981). **The testing trap.** New York: Rawson Wade Publishing.

Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L., Merriman, H. O., & Provus, M. M. (1971). **Educational evaluation and decision-making.** Itasca, IL: Peacock.

Stufflebeam, D. L., & Shinkfield, A. J. (1985). **Systematic evaluation.** Boston: Kluwer-Nijhoff.

Swaminathan, H. (in press). Bayesian estimation in the two-parameter logistic model. **Psychometrika.**

Swaminathan, H., & Gifford, J. (1983). Estimation of parameters in the three-parameter latent trait model. In C. Weiss (Ed.), **New horizons in testing: Latent trait theory and computerized adaptive testing.** New York: Academic Press.

Taba, H. (1962). **Curriculum development.** New York: Harcourt, Brace, & World.

Tinerow, M. M. (1984, November). **Traditional and nontraditional educational elements using telecommunications.** Paper presented at the National Adult Education Conference, Louisville. (ERIC Document Reproduction Service No. ED 249 361)

Tittle, C. K. (1984, April). **Professional standards and equity: The role of evaluators and researchers.** Paper presented at the meeting of the American Educational Research Association, New Orleans.

Torshen, K. (1969). **The relation of classroom evaluation to students' self-concepts and mental health.** Unpublished doctoral dissertation, University of Chicago.

Travers, R. M. W. (1983). **How research has changed American schools.** Kalamazoo, MI: Mythos Press.

Turkle, S. (1984). **The second self: Computers and the human spirit.** New York: Simon & Schuster.

Twentieth Century Fund, Task Force on Federal Elementary and Secondary Education Policy. (1983). **Making the grade.** New York: The Twentieth Century Fund.

Tyler, R. W. (1968). Critique of the issue on educational and psychological tests. **Review of Educational Research**, 38, 102-107.

Tyler, R. W. & White, S. W. (1979). **Testing, teaching, and learning.** Report of a Conference on Research on Testing. Washington, D.C.: National Institute of Education, U.S. Department of Health, Education and Welfare.

Weber, G. (1974). **Uses and abuses of standardized testing in the schools.** (Occasional paper No. 22). Washington, D.C.: Council for Basic Education.

Weckstein, P. (1973). **Legal challenges to educational testing practices.** Harvard Center for Law and Education, Classification Materials, 186-198.

Weckstein, P. (1976). **Legal challenges to educational testing practices** (Supplement). Harvard Center for Law and Education, Classification Materials, 37-38.

West, T. W. (1984). The development of a network of telecommunications networks: A contagion period. **Cause/Effect,** 7 (5), 2-3.

Williams, R. L. (1971). Abuses and misuses in testing black children. **Counseling Psychologist,** 2, 62-67.

Willis, S. (1972). **Formation of teachers' expectations of students' academic performance.** Unpublished doctoral dissertation, University of Texas at Austin.

Wolf, R. L. (1979). The use of judicial evaluation methods in the formulation of educational policy. **Educational Evaluation and Policy Analysis,** 1 (3), 19-28.

Wright, B. D., & Bell, S. R. (1984). Item banks: What, why, and how. **Journal of Educational Measurement,** 21, 331-345.

APPENDIX 1

Personnel Distribution Table

Personnel Distribution Table

Category	Minorities	Women	Directors	Core Staff	Consultant
Peter Airasian				x	
Cordelia Alexander	x	x			x
Gilbert Austin				x	
Mary Anne Bunda		x		x	
Judith Burry		x			x
William Cooley				x	
William Denton				x	
Esther Diamond		x			x
Susan Embretson		x			x
Arnold Gallegos	x				x
Douglas Glassnap				x	
Maria Luisa Gonzalez	x	x			x
Robert Grobe					x
Egon Guba				x	
Ron Hambleton			x		
Walter Haney				x	
Thomas J. Heiry					x
Grace Iverson		x		x	
Richard Jaeger					x
Yvonna Lincoln	x	x		x	
Larry Ludlow					x
George Madaus			x		
Samuel Mayo					x
Robert Mendro				x	
Jack Merwin				x	
Edward Meyen					x
M. David Miller					x
Napoleon Mitchell	x			x	
George H. Olson					x
Joseph Pedulla				x	
John Poggio			x		
Sri Kanta Rao	x				x
Thomas Ryan					x
John Sandberg					x
James Sanders			x		
Michael Scriven					x
Daniel Stufflebeam			x		
Hariharan Swaminathan	x			x	
Lyke Thompson			x		
Carol Kehr Tittle		x		x	
Robert Travers					x
Ralph W. Tyler				x	
Richard Wallace				x	
William Webster			x		
Pecent of Staff	**13%**	**20%**	**16%**	**41%**	**43%**

APPENDIX 2

Chart of Cooperating Agencies

	BOSTON COLLEGE	UNIV. KANSAS	U.M.U.	DALLAS SCHOOLS
School Dists.	City of Boston School Committee E.Greenwich,RI Hudson, NH Lynnfield, MA Middleborough MA New Bedford, MA Newport, RI N.Brookfield,MA Pawtucket, RI Sanborn,Kingston Newton, RI Silver Lake, Kingston, MA We have letters from 31 school districts in Massachusetts, New Hampshire, & Rhode Island which have committed themselves to collabortion with the Center participating in the state minimum competency testing program	Kansas Dists.[3] Jefferson County Kentucky Waterloo, IA[1]	Indianapolis[3] Kalamazoo[1] Cincinnatti, OH[1] Saginaw, MI[1] Comstock, MI Hillsborough, County, FA Detroit Lansing, MI[1] Springfield, OR[1] Shaker Hts., OH[1] Toledo[3]	Texas Consortium[1] (Eight largest urban schools districts in Texas Fort Worth Atlanta
Labs		MCREL[1]	NWREL[2] Midwest Lab[1] SWREL[1] AEL[1]	
Centers	LRDC[1] CSOS[1]		IRT[2] NCVTE[1] Oregon Ctr.[1] (CEPM)	UTTEC[1]

	BOSTON COLLEGE	UNIV. KANSAS	W.M.U.	DALLAS SCHOOLS
Prof. Society	CAPE[1] NCME[1] NPTA[1] APA[1] ASCD[1] AFT[1]	ACE[1]	Joint Comm.[2] Black Coll.2 EN[1] NEA[1] AASA[1] ECS[1] APGA[1] NASSP[1] NAESP[1] NSBA[1]	AERA Div. A[1]
Govt. Progs.	ERIC T/M[3] JDRP/NDN[1]			
EAs	Mass1 South Carolina	Kansas[1]	Maryland[3] Michigan[1] Minnesota[1] ECS[1] LA.[3]	TEA[1]
Individuals and Others	Ron Hambleton[4] Hariharan Swaminathan[4] Peter Airasian[4] Walter Haney[4] Lary Ludlow[1] George Madaus[4]	Yvonne Lincoln[3] Judith Burry[3] Susan Embretson[1] Douglas Glasnapp[4] Egon Guba[4]	Mike Scriven[3] Daniel L. Stufflebeam[4] Egon Guba[3] Gil Austin[4] Mary Anne Bunda[4] Arnold Gallegos[4]	Cordelial Alexander[3] Bill Denton[3] Maria Luisa Gonzalez[1] Robert Grobe[3] Thomas Heiry[1]

	BOSTON COLLEGE	UNIV. KANSAS	W.M.U.	DALLAS SCHOOLS
Individuals and Others	Joseph Paolulla[4] Gil Austin[3]	Ed Meyen[1] David Miller[3] John Poggio[4]	Thomas Ryan[1] Carol Tittle[1] Sri Kanta Rao[1] Bob Travers[3] Dick Jaeger[3] Lyke Thompson[4] Bill Cooley[4] E. Diamond[1] John Sandberg[3] James Sanders[4] Robert Travers[3] S. Mayo[1] R. Tyler[4]	Robert Mendro[4] Napoleon Mitchell[4] George Olson[1] Bill Webster[4]

	BOSTON COLLEGE	UNIV. KANSAS	W.M.U.	DALLAS SCHOOLS
Non Ed Groups External	Consumers Union			

1. Communication
2. Coordination
3. Cooperation
4. Collaborator

APPENDIX 3

Key Informant Survey Summary

Key Informant Survey Summary

Respondent	Priorities and Recommendations
National Urban League Ms. Stephanie Robinson Director of Education	Validating new approaches and programs Documenting effective programs Disseminating information replicating effective programs
Center for Law and Education Ms. Sue Jackson Board Member	Develop alternatives to traditional multiple choice tests that reinforce critical thinking, problem solving, and effective writing and speaking Develop and catalogue existing testing devices that are fair and equitable and relate to different learning styles and areas Set a series of goals and objectives for testing and its effects and work toward them Establish models of parent-student-staff evaluations and utilizations of evaluations Develop more effective ways of describing educational outcomes
Council for Basic Education Mr. Dennis Gray Deputy Director	Educating lay audiences about proper uses of testing; including parents, policymakers, journalists, and other citizens Linking testing to curriculum and instruction Teaching teachers how to use testing properly Moving beyond paper and pencil testing to performance reviews Testing programs for "higher order" thinking
American Association of School Administrators Dr. Richard D. Miller Executive Director	Test validation Dissemination of results
Mexican American Legal Defense Fund Mr. Ron Vera Attorney & Director Higher Education Project	Procedures for verifying test scores on standardized tests Strengthening the capability of administrators to utilize and interpret standardized tests Validating the policy uses to which standardized tests are put
Southwest Educational Development Laboratory Dr. Preston Kronkosky Executive Director	Synthesize and disseminate the existing knowledge bases on educational testing Serve as a clearinghouse for collecting, storing, retrieving, and disseminating information on school testing evaluation and standards

	Develop alternatives to the traditional multiple choice tests that better reinforce critical thinking, problem solving, and effective writing and speaking Designing procedures for testing and evaluation to provide systematic and fair placement and promotion of students Setting appropriate standards guidelines against which student test performance may be compared
Public Education Association Ms. Jeanne Frankl Executive Director	Development of tests that; focus on thinking, assess developmental levels, are presciptive/-diagnostic, teacher useable and manageable, and measure growth in skill Focus on appropriate utilization of tests for promotion, on setting standards for teacher made tests(most closely allied to classroom work), and testing of professional, as well as, student competencies Long range study of the predictive value of test results Development of multi-racial, multi-cultural non-verbal test to serve all LEP children Study the equity implications of private coaching schools for tests like SAT
University of North Dakota Center for Teaching and Learning Dr. Vito Perrone Dean	Study of equity in relation to testing, testing and curriculum related to teaching and learning, alternative assessment mechanisms
The James Russell Lowell School Dr. William D. Corbett Principal	Examine the effects on education in general and student writing in particular of the intense use of multiple choice, fill in testing Explore alternate methods of testing Establish openness in testing at all levels Establish ERIC test dissemination contract Study the impact of test-directed placement policy on special populations
National Council of English Teachers Dr. Yetta Goodman President	Develop testing and evaluation techniques that facilitate a range of curriculum models Study of truth in testing and of legality in relation to testing Study the impact of testing on minority populations
Michigan-Elementary and Middle Schools Principals Association Dr. William Mays, Jr. Executive Director	Serve as a clearinghouse for disseminating educational research Study of test and evaluation procedures for the fair placement and promotion of students Effects of linking school funding to test score and of linking promotion/pay raises to test scores Study policy making process by which standards for assessing students are set

National Science Teachers Association Dr. Edward P. Ortleb President	Develop alternatives to the traditional multiple choice tests that better reinforce critical thinking, problem solving, and effective writing and speaking Develop evaluation techniques applicable to local instructional programs Link testing more closely to instruction Study positive and negative effects of developing educational policy on tests result Develop methods to aid teachers and school administrators to interpret test data Develop appropriate guidelines for setting standards against which student test performance may be compared Study the policy making process by which standards for assessing students are set
National Education Association Dr. Bernard McKenna Program Development Specialist	Evaluation of student learning in the broad sense Development of multiple criteria for evaluating student progress Develop assessment for diagnosis, prescribing remediation, and planning instruction in general
Virginia-Department of Education Dr. Gerald W. Bracey Special Assistant for Policy	Development of proficiency based assessments Study the appropriate role of assessments in policy Study the current and desired relation of instruction to assessment Develop long-term strategies for data base Study of norm-referenced testing in relation to their construction
Kansas-National Education Association Dr. Marilyn Flannigan Diretor of Instructional Advocacy	Develop more effective ways of describing educational outcomes Develop testing practices which assist vulnerable student populations
Association for Supervision and Curriculum Development Dr. Gordon Cawelti	Study discrepencies of national norm-referenced tests and local curriculum Help districts develop subject matter lists on their curriculum as an alternative to reliance and preoccupation with SAT or NRT as indicators of excellence Study the effects of testing not specifically linked to content, such as "gate testing", which is responsible for promotion, advancement, placement, or acceptance Develop methods for presenting test results(norm or criterion referenced) to parents and other lay audiences in meaningful and useful manners Develop valid and reliable standards for passing, or "cut points"

APPENDIX 4

Content Analysis of Key Informant Survey

NIE Planning Survey: Responses to Open-Ended Questions

Key

Dimension A

#	Code
53	1: Testing
6	2: Evaluation
0	3: Standards
2	4: Testing & Evaluation
4	5: Testing & Standards
0	6: Evaluation & Stndrds
1	7: All Three
2	8: None of the Above
68	

Dimension B

#	Code
10	1: Enhance curriculum & instruction
5	2: Promote diagnostic/prescriptive uses
8	3: Disseminate information to constituents
21	4: Enhance equity/reduce inequity
4	5: Enhance planning/policy uses
13	6: Reduce problems with traditional approaches/promote uses of alternative approaches
7	7: Promote student growth/higher order learning
68	

Code A	Code B	Comment
Section A: MISSION GOALS/OBJECTIVES FOR THE R & D CENTER		
1	3	Develop public information programs on limitations of standardized tests.
1	3	Study & disseminate information on the impact of minimal competency testing programs on minority education.
1	4	Development of multi-racial, multi-cultural nonverbal test to serve all LEP children.
1	4	Establish the practice of test openness in all publically supported or required testing.
1	4	See Golden Rule settlement (enclosed openness in testing article?).
1	4	Investigate testing abuse.
1	6	Long range study of predictive value of test results.
2	6	Develop in process methods of evaluations.
2	7	Multiple criteria for evaluating student learning progress.
5	4	Develop standards for truth in testing.
Section B: INTEGRATION OF TESTING WITH STUDENT LEARNING		
1	1	How to keep testing from dominating curriculum.
1	4	Examination of the discriminatory features of tests for specific populations including boys for reading in early grades, women, multiligual populations, etc.

Code A	Code B	Comment
8	4	Issues related to cultural bias.
Section C: ETHICAL AND POLITICAL ISSUES		
1	4	Equity implications of private coaching schools for tests like SAT (see item C4, the effects and implications of testwiseness).
1	4	When tests are used to judge a person, that person should have the absolute right to view the test and his/her scored answers post administratively.
1	4	Sould tests be used for placement and promotion? Can any one test be a gatekeeper?
1	4	The use of tests as the major means of discrimination in our society.
1	4	The use of tests to screen out othrewise qualified people from programs, jobs, education, professions, etc.
Section D: TEST INFORMATION UTILIZATION		
1	6	Informal observational techniques in testing (Kidiva & Ching).
1	6	Limitation of test-teach-test paradigm.
2	1	Evaluation (observational techniques) as a continuous part of curriculum development and instruction.
Section E: STANDARDS-SETTING		
1	2	Testing assessment for diagnosis.
1	2	Testing assessment for prescribing remediation.
5	4	What are the purposes for setting standards? Everything involving standards encorporates a view of test as "gatekeeper." This issue itself needs to be examined in terms of its purposes relevant to society and various groups within the society.
Section G: MOST CRITICAL NEEDS		
1	1	Linking testing to curriculum and instruction.
1	1	Testing & curriculum/teaching & learning.
1	1	Major problem is incongruity of national norm referenced tests and and local curriculum--need more work similar to IRT pointing out descrepancies.
1	1	Helping instructors develop subject matter tests on their curriculum vs. prescribing with SAT or NRT as indicators of excellence.
1	1	Assessment for planning instruction generally.

Code A	Code B	Comment
1	1	Relation of instruction to assessment (current & desired).
1	2	Prescriptive/diagnostic tests.
1	2	Assessment for diagnosis.
1	2	Assessment for remediation.
1	3	Educating lay audiences about proper uses of testing: parents, policy makers, journalists, other citizens
1	3	Teaching teachers how to use testing properly.
1	3	Dissemination of results.
1	3	A1. Synthesize & disseminate existing knowledge bases on educational testing.
1	3	Description of how NRT's are really constructed.
1	4	Equity in relation to testing.
1	4	Establishing openness in testing at all levels.
1	4	Truth in testing.
1	4	Legality and testing.
1	5	Appropriate role of assessment in policy.
1	6	Moving beyond paper and pencil testing to performance reviews.
1	6	Test validation.
1	6	A8. Develop alternatives to the traditional multiple choice tests that better reinforce critical thinking, problem solving, and effective writing and speaking.
1	6	Teacher useful and manageable tests.
1	6	Alternative assessment mechanisms.
1	6	Misuse of standardized entrance tests by colleges and universities throughtout the country. (See included document.)
1	6	Examining the effects on education in general and student writing in particular of the intense use of multiple choice, fill in testing.
1	6	Exploring alternative methods of testing.
1	6	Need for proficiency-based assessment.
1	7	Testing programs for "higher order" thinking.

Code A	Code B	Comment
1	7	Tests that focus on thinking competencies.
1	7	Tests assessing developmental levels.
1	7	Tests that measure growth in skill.
2	4	Impact of evaluation on minority populations.
2	7	Evaluation of student learning in the broad sense.
2	7	Multiple criteria for evaluating student progress.
4	1	Testing and other evaluation techniques that facilitate a range of curriculum models.
4	5	C3. Testing and evaluation procedures for systematic and fair placement and promotion of students.
5	5	E1. Appropriate guidelines for setting standards against which student test performance may be compared.
7	3	A2. Serve as a clearing house for collecting, storing, retrieving, and disseminating information on school testing, evaluation, and standards.
8	5	Need for strategic thinking about data (as opposed to quick fix remedies based on short term information).
Section H: ADDITIONAL RECOMMENDATIONS		
1	4	Focus on appropriate utilization of test results for promotion, etc.
1	4	I assume they will focus on testing of professional competencies as well as student competencies.
1	4	I admire your effort to establish a national testing center. It is long overdue. The center should be staffed with a wide variety of professionals including knowledgable practitioners at all levels of the educational spectrum.
5	1	Focus on how to set standards for teacher-made tests--most closely allied to classroom work.

APPENDIX 5

Content Analysis of Summary Report of
Reviewer's Comments-Planning Grant Competition

NIE PLANNING GRANTS COMPETITION -- SUMMARY REPORT OF REVIEWERS' COMMENTS

Summary of Reviewers' Judgments of Proposal Components in the Form of <Reviewer> judges the <component> as <judgment>.

<Reviewer> judges	the <component>	as <judgment>.
Introduction		
1) The panel	application	mean rank of 2.4.
Major Strengths		
1) A majority of the panelists	proposal on each of the evaluation criteria	generally strong.
2) Each of the panelists	summary of the history of developments in testing & evaluation	strength.
3) Each of the panelists	analysis of the interrelatedness of testing, evaluation, & standards	strength.
4) ---	developing a conceptual framework related to the Center mission	good job.
5) ---	plan of operation	well structured, appropriate for meeting stated objectives of planning process.
6) ---	organizational plan in general	exceptionally strong.
7) ---	strategies for outreach in particular	exceptionally strong.
8) Several panel members	applicant's concern & plans for addressing the issues of the underserved	strength.
9) ---	overall quality of proposed staff	major strength.
10) ---	(staff) in psychometric aspects of testing	somewhat lacking.
11) The panel	(staff) Stufflebeam & Medaus	noted accomplishments of.

[reviewer] judges	the <component>	as <judgment>.
12) ---	(staff) proposed consultants	"breadth of experience."
13) ---	(staff) the group	"good mix of scholars & practitioners."
14) ---	evidence of prior collaboration among institutions	strength.
15) ---	"track record" of prior institutional accomplishments	strength.
Major Weaknesses		
1) Each panel member	mission statement	the primary concern--overall vagueness and lack of specificity.
2) Panel members	important issues in the field & how these issues might be addressed	lacking sufficient specificity.
3) ---	psychometric and and other technical issues regarding testing	far too little attention.
4) ---	proposal	somewhat imbalanced toward evaluation to the neglect of important testing concerns.
5) ---	scope of the plan of operation	much too extensive.
6) The panel	extensive schedule of interviewing	might lead to difficulty in consolidating the obtained input.
7) ---	design	seems to delay discussion of research priorities . . . thus compresses the hard choices into a "painfully narrow time span,"
8) The panel	inclusion of teacher evaluation as a focus of the mission statement	questioned its appropriateness for a center on "student" testing, evaluation & standards.

<Reviewer> judges	the <component>	as <judgment>.
Summary of Analysis		
1) The panel majority	proposal	a strong application that is fundable.
2) Only one of the panel members	proposal	"fundable with reservations."
3) ---	plan of operation	particularly strong.
4) ---	strategies for outreach	particularly strong.
5) ---	what can be accomplished within a particularly short time frame	perhaps overly ambitious.
6) ---	discussion of critical technical and psychometric issues	somewhat lacking.
7) ---	mission statement	primary criticism--vague and overly general giving little basis for judging the direction the Center effort might take.
8) ---	operating plan as a vehicle for determining the Center's agenda	the strength that overcomes the above weakness.

APPENDIX 6

Content Analysis of Request for Proposal
for NIE Center on Student Testing,
Evaluation, and Standards Setting

SUMMARY OF 1985 RFP FOR NIE CENTER ON STUDENT TESTING, EVALUATION, AND STANDARDS
LIST OF ONLY THOSE STATEMENTS DIRECTLY RELATED TO TESTING

TOPIC
T E S where T = TESTING, E = EVALUATION, S = STANDARDS.

Definitions

X **Testing:** Formal means to assess student educational outcomes. This may refer to paper & pencil testing, performance testing, and other forms of testing.

X X **Assessment:** Not explicitly defined. Usually implies "testing," other times implies "testing-based evaluation."

Introduction

Context

X X 1. Testing & evaluation activities play a central role in American schools.

X X 2. Parents, educators, & policymakers at state, national, & local levels ask questions about a) how their children plus specific schools & schools in general are doing, & b) how can they be improved.

X X 3. Testing & evaluation activities can answer many of these questions.

X X 5. Policymakers & legislators often act upon testing & evaluation results.

X X X 6. The perception of crisis in current educational standards is largely based upon results from testing & evaluation.

Strengths/Opportunities

X 1. Test developers & publishers have produced many standardized tests for large scale use in American education, e.g., norm-referenced tests, the SAT, & other test of aptitude & capacity for post secondary education.

X X 2. The National Assessment of Educational Progress has provided periodic assessments of the nation's overall progress in education for the past five years.

X X 3. State assessments of different kinds have provided much information for educational decision making at state & local levels.

X X 4. Progress is being made in the area of local testing & evaluation because of a) advances in the broader arena—state & national, & b) developments in learning theory & technology, testing & evaluation.

TOPIC
T E S where T = TESTING, E = EVALUATION, S = STANDARDS.

Needs/Concerns/Problems

X X 1. The need for better testing & evaluation continues.

X X 4. A need for better testing & evaluation tools already exist locally.

X X 5. The benefits attained in developing testing & evaluation methodologies for state & national use have not always been transferable to the local level.

X X 6. Concerns at the local level frequently focus on the fairness & utility of testing & evaluation methods for dealing with individual students, classrooms, & other groupings.

X X 7. A need for more readily available & systematic information based on testing & evaluation to help diagnose, place, instruct, & promote individual students exists.

X X 8. A need exists for tools to assess the merits of individual classroom & school programs more satisfactorily because statewide or norm-referenced test results are often the only ones available with which to evaluate them. Such tests are not good for assessing the effectiveness of discrete local instructional activities. This need will grow as school districts are called upon to implement comprehensive reform legislation initiated by many states.

X X 9. A need still exists to identify ways to make testing & evaluation information more useful to its intended audiences.

X 10. Part of the challenge is technical, e.g., to present psychometric information clearly & succinctly or to package & distribute information in timely & efficient ways.

X X 11. Part of the challenge entails better translation of teachers' & other users' language to that of testing & evaluation & visa versa.

X X 12. It also means integrating the perspectives of teachers & others into testing & evaluation so that the different components of school work more harmoniously together.

X X 13. The amount of attention & resources placed upon local testing & evaluation has historically lagged behind that accorded to the issues beyond the local level.

Scope of Center Mission

Specifications: Mission/Shoulds/Mays

X X 1. The primary mission of this Center should be to increase the contribution that testing & evaluation can make to local school improvement.

X X 3. The Center ought to make demonstrable contributions toward meeting two goals: 1) a rise in test scores, and 2) a perception that educational standards have improved. Examples of such contributions include:

X X * developing more effective, efficient, & fair methods of testing & evaluation in schools;

X X * demonstrating ways to make testing & evaluation practices more equitable to the individual students; &

X X * demonstrating ways to make these practices more useful to all concerned at the local level.

X X 4. The Center should be cognizant of the impacts of its research upon policy at the state & national level & maintain contacts with officials at these levels, including state research & evaluation directors & state assessment directors.

X X 6. Testing & evaluation in the core subjects of reading, language arts, writing, & mathematics should be emphasized. The mix of these subjects is optional.

X X 7. Research on testing & evaluation in other subjects such as social studies, the arts & humanities, or nonachievement outcomes like attendance or dropout rates may be addressed if it can be related to other work on the core subjects, or exceptional expertise or unusual opportunity presents itself.

X X 8. The Center should address testing & evaluation issues in elementary & secondary education.

X 9. Issues at the secondary level may include the use of proficiency, minimum competency, or other achievement tests used to regulate high school graduation.

Strengths

X X X 1. Good testing & evaluation can provide information with sufficient precision, timeliness, & utility to assist in the debate & decisionmaking about standards.

Research That May be Performed

General Specifications: Shoulds/Mays

X X 2. Basic research should be framed so that its potential value to testing & evaluation in schools can clearly be seen.

X 3. Research may embody theoretical perspectives of classical test design, item response theory, Bayesian statistics, cognitive science & psychology, subject matter domains & technology.

X X 4. Research may involve practical perspectives of testing & evaluation procedures & use.

X X 5. Research may involve both laboratory & field work, but a substantial proportion of the research should be conducted on the school sites & involve the active participation of school personnel such that a reasonable observer ought to be able to conclude that the extent of research conducted in schools is commensurate with the Center's mission to help school improvement through testing & evaluation.

Technological Applications

1. **Diagnostic Testing**

X **Needs:** a) A need for efficient & perceptive forms of diagnostic testing for use with individuals or smaller groups of students exists.

X **Strengths:** a) Progress in diagnostic testing has been aided by advances in microcomputer technology, cognitive science, & psychometric theory.

X **Questions:** a) How can specifications for such tests be best written, particularly regarding content?

X b) How short can such tests be and still be reliable and valid measures of student learning?

2. **Tailored or Adaptive Testing**

X **Strengths:** a) Such tests offer the possibility of reducing testing time, thereby making classroom testing more efficient.

X b) Such testing adapts the starting point and questions to be asked to the ability level of the individual student.

X c) A considerable body of theoretical knowledge of such tests exists.

X X **Needs:** a) Further research to establish the validity as well as the efficiency and utility of such testing might be conducted.

3. **Nonintrusive Testing**, in which instructional software is designed to ascertain what the student is learning by monitoring the level and sequence of materials attempted.

X **Strengths:** a) The student is tested without taking a "test."
b) Capacities in technological hardware and software are increasing.

X c) Learning and psychometric theory are advancing.

X X **Needs:** a) Better testing & evaluation in schools is needed.

Locally Responsive Tests

1. **Tests That Measure the Specific Objectives of Local Instructional Programs**

X **Needs/Problems:** a) An increasing need for such tests exists.

X b) A good measure of a local reading program at the intermediate grades may be needed.

X c) A conventional norm-referenced test may be neither matched with the content of the local program nor conclusive as to what knowledge the students actually learned.

X d) Large numbers of subject domains have to be assessed, making the local cost of developing tests for all of them very high and the workload excessive.

X **Strengths:** a) The development of criterion-referenced tests, which assess mastery of specific subject domains, has helped meet school needs in this area.

2. **Item Banks**

X **Strengths:** a) Item response theory has increased the capacity to generate banks of test items that may be used interchangeably by school districts and state education agencies to create customized tests for their own use.

X b) Item response theory makes possible the definition of the statistical properties of individual test items without reference to the rest of the test with which they were developed.

X c) Alternatives to item banks may also meet the need for locally responsive tests.

T E S where T = TESTING, E = EVALUATION, S = STANDARDS.

X **Questions:** a) How can test specifications best be developed for local use?
X b) Under what conditions is item banking a productive test development strategy for local use and how can those conditions be maximized?
X c) When are other methods more appropriate?
X **Needs:** a) Schools and districts need locally responsive tests that are not now available.

Curriculum-Test Match

X **Needs:** a) Research might further clarify the consequences of the relationship between the curriculum in school districts and the tests used in those districts.
X **Questions:** a) When tests indicate that students have not mastered a subject, to what degree are the tests an accurate measure of what was taught?
X X Strengths: a) Further research in this area & the others mentioned above would lead to more conclusive measures of how well students are doing and what present educational standards are.

Helping with Educational Standards

X X X **Needs/Problems:** a) Current efforts to raise educational standards involve many aspects of testing & evaluation whose methodology is imperfect or undeveloped, e.g., the use of proficiency tests, etc., to regulate promotion and graduation at the local level requires instruments that are fair to all students, particularly lower-achieving ones and those near the threshold which has been established as the "passing" or "cut" point.
X X **Questions:** a) How can subject matter domains be better specified and tests made more discriminating at the critical "cut points"?
X X b) What local nonpsychometric procedures (e.g., methods for standard setting) can be developed to make pass-fail decisions more accurate and equitable?

Assisting Teachers and Parents with Testing and Evaluation

X X **Needs/Problems:** a) Many practical methodological concerns exist at the local level about the capacity of teachers, other staff, & parents to use testing & evaluation for day-to-day purposes.
X X b) Teachers & parents need means to help them understand effective testing & evaluation practices and adapt such practices to the needs of classrooms.
X X **Questions:** a) How can testing & evaluation results best be made responsive to the needs of teachers and other local staff members?
X b) What degree of translation between goals of teachers and those of test developers is necessary to turn teachers' instructional objectives into useful tests and test items?
X X c) What are the best strategies to engage teachers, other local staff members, and parents in testing & evaluation methods?
X X d) What are the limits to such strategies?

TOPIC
T E S where T = TESTING, E = EVALUATION, S = STANDARDS.

Information Systems

Strengths/Context:

X X b) Such systems may contain results of testing programs, and other evaluative data, such as statistics on class grades, homework, disciplinary measures, attendance, and dropouts.

Evaluation and Testing Uses: How to Gain More Use of Credible Information

X X **Problems/Context:** a) People in all quarters of education display varying degrees of reluctance to use good testing & evaluation.

X X c) In some cases, people have been "burned" by the use of testing & evaluation information that actually turned out to be faulty in one or more aspects.

X X d) The research topics suggested above address the problem of faulty or technically deficient information through technical or applied perspectives while this topic emphasizes better understanding of how to gain more use of credible testing & evaluation information.

X X **Questions:** a) What barriers exist to the use of credible testing & evaluation information?

X X b) How can the discovered barriers be overcome?

X X c) How can the production of good information be made less costly?

X X **Shoulds:** a) Research in this area should help establish what the potential of the "information age" at the local school level actually is.

X X b) Research should not address general topics such as what distinguishes the school as an information-seeking institution.

X X c) Any studies of information use should be tightly tied to discrete issues of testing and evaluation practice.

Organization and Staffing

Specifications: Shoulds/Mays

X X 5. Staff members with practical perspectives on testing & evaluation procedures based on experience in schools are explicitly desired as part of the Center's overall staffing.

Dissemination

Specifications: Shoulds/Wills

X X 2. The Center will make its contribution to scholarly journals in testing & evaluation.

X X X 4. It will serve as a resource to those in schools & elsewhere concerned with improving schools in the areas of testing, evaluation, and standards, e.g., school district directors of research & evaluation, directors of testing, teachers & teacher organizations, state research & evaluation directors, and state assessment directors.

X 5. The Center's work will be made widely available to educational test publishers and similar organizations.

TOPIC T E S	where T = TESTING, E = EVALUATION, S = STANDARDS.
X X	8. Cooperation with the ERIC Clearinghouse on Tests, Measurement, and Evaluation is expected.

Collaboration with Other Centers

Context/Needs/Strengths

X X X	3. The NIE Center on Testing, Evaluation and Standards will have special skills that may be selectively used on challenges of the other Centers, e.g., it might allocate some staff time & resources to work collaboratively with another Center on a problem of mutual interest. Ideally, this would happen where both significant testing & evaluation issues and significant substantive issues were raised in a particular research problem. Other situations can also be imagined.

Examples

X X	1. Writing (writing assessment in the field)
X X	2. Reading (reading assessment in the field) not part of this grants competition
X	6. Postsecondary Teaching and Learning (basic skills or other testing used for admission or placement of students in postsecondary institutions)
X X	7. Teacher Quality and Effectiveness (methods for assessing teacher competencies)

SUMMARY OF 1985 RFP FOR NIE CENTER ON STUDENT TESTING, EVALUATION, AND STANDARDS
LIST OF ONLY THOSE STATEMENTS DIRECTLY RELATED TO EVALUATION

TOPIC
T E S where T = TESTING, E = EVALUATION, S = STANDARDS.

Definitions

X **Evaluation:** Process used to place student educational outcomes in their specific context, relate them to instructional and other factors, weigh their significance, and make decisions based on them.

X X **Assessment:** Not explicitly defined. Usually implies "testing," other times implies "testing-based evaluation."

Introduction

Context

X X 1. Testing & evaluation activities play a central role in American schools.

X X 2. Parents, educators, & policymakers at state, national, & local levels ask questions about a) how their children plus specific schools & schools in general are doing, & b) how can they be improved.

X X 3. Testing & evaluation activities can answer many of these questions.

X 4. School personnel are held accountable for their results.

X X 5. Policymakers & legislators often act upon testing & evaluation results.

X X X 6. The perception of crisis in current educational standards is largely based upon results from testing & evaluation.

Strengths/Opportunities

X X 2. The National Assessment of Educational Progress has provided periodic assessments of the nation's overall progress in education for the past five years.

X X 3. State assessments of different kinds have provided much information for educational decision making at state & local levels.

X X 4. Progress is being made in the area of local testing & evaluation because of a) advances in the broader arena—state & national, & b) developments in learning theory & technology, testing & evaluation.

Needs/Concerns/Problems

X X 1. The need for better testing & evaluation continues.

X X 4. A need for better testing & evaluation tools already exists locally.

X X 5. The benefits attained in developing testing & evaluation methodologies for state & national use have not always been transferable to the local level.

X X 6. Concerns at the local level frequently focus on the fairness & utility of testing & evaluation methods for dealing with individual students, classrooms, & other groupings.

X X 7. A need for more readily available & systematic information based on testing & evaluation to help diagnose, place, instruct, & promote individual students exists.

X X 8. A need exists for tools to assess the merits of individual classroom & school programs more satisfactorily because statewide or norm-referenced test results are often the only ones available with which to evaluate them. Such tests are not good for assessing the effectiveness of discrete local instructional activities. This need will grow as school districts are called upon to implement comprehensive reform legislation initiated by many states.

X X 9. A need still exists to identify ways to make testing & evaluation information more useful to the intended audiences.

X X 11. Part of the challenge entails better translation of teachers' & other users' language to that of testing & evaluation & visa versa.

X X 12. It also means integrating the perspectives of teachers & others into testing & evaluation so that the different components of school work more harmoniously together.

X X 13. The amount of attention & resources placed upon local testing & evaluation has historically lagged behind that accorded to the issues beyond the local level.

Scope of Center Mission

Specifications: Mission/Shoulds/Mays

X X 1. The primary mission of this Center should be to increase the contribution that testing & evaluation can make to local school improvement.

X X 3. The Center ought to make demonstrable contributions toward meeting two goals: 1) a rise in test scores, and 2) a perception that educational standards have improved. Examples of such contributions include:

X X * developing more effective, efficient, & fair methods of testing & evaluation in schools;

X X * demonstrating ways to make testing & evaluation practices more equitable to the individual students; &

X X * demonstrating ways to make these practices more useful to all concerned at the local level.

TOPIC
T E S where T = TESTING, E = EVALUATION, S = STANDARDS.

X X 4. The Center should be cognizant of the impacts of its research upon policy at the state & national level & maintain contacts with officials at these levels, including state research & evaluation directors & state assessment directors.

X X 6. Testing & evaluation in the core subjects of reading, language arts, writing, & mathematics should be emphasized. The mix of these subjects is optional.

X X 7. Research on testing & evaluation in other subjects such as social studies, the arts & humanities, or nonachievement outcomes like attendance or dropout rates may be addressed if it can be related to other work on the core subjects, or exceptional expertise or unusual opportunity presents itself.

X X 8. The Center should address testing & evaluation issues in elementary & secondary education.

Strengths

X X X 1. Good testing & evaluation can provide information with sufficient precision, timeliness, & utility to assist in the debate & decisionmaking about standards.

Research That May be Performed

General Specifications: Shoulds/Mays

X X 2. Basic research should be framed so that its potential value to testing & evaluation in schools can clearly be seen.

X X 4. Research may involve practical perspectives of testing & evaluation procedures & use.

X X 5. Research may involve both laboratory & field work, but a substantial proportion of the research should be conducted on the school sites & involve the active participation of school personnel such that a reasonable observer ought to be able to conclude that the extent of research conducted in schools is commensurate with the Center's mission to help school improvement through testing & evaluation.

Technological Applications

1. **Diagnostic Testing**
   Questions:
X c) What kinds of diagnostic information have practical value for the classroom teacher and other school staff?

2. **Tailored or Adaptive Testing**
X X **Needs:** a) Further research to establish the validity as well as the efficiency and utility of such testing might be conducted.

TOPIC
T E S where T = TESTING, E = EVALUATION, S = STANDARDS.

---

3. **Nonintrusive Testing**, in which instructional software is designed to ascertain what the student is learning by monitoring the level and sequence of materials attempted.

X X **Needs:** a) Better testing & evaluation in schools is needed.

Helping with Educational Standards

X X X **Needs/Problems:** a) Current efforts to raise educational standards involve many aspects of testing & evaluation whose methodology is imperfect or undeveloped, e.g., the use of proficiency tests, etc., to regulate promotion and graduation at the local level requires instruments that are fair to all students, particularly lower-achieving ones and those near the threshold which has been established as the "passing" or "cut" point.

Assisting Teachers and Parents with Testing and Evaluation

X X **Needs/Problems:** a) Many practical methodological concerns exist at the local level about the capacity of teachers, other staff, & parents to use testing & evaluation for day-to-day purposes.

X X b) Teachers & parents need means to help them understand effective testing & evaluation practices and adapt such practices to the needs of classrooms.

X X **Questions:** a) How can testing & evaluation results best be made responsive to the needs of teachers and other local staff members?

X X c) What are the best strategies to engage teachers, other local staff members, and parents in testing & evaluation methods?

X X d) What are the limits to such strategies?

Information Systems

Strengths/Context:

X X b) Such systems may contain results of testing programs, and other evaluative data, such as statistics on class grades, homework, disciplinary measures, attendance, and dropouts.

Evaluation and Testing Uses: How to Gain More Use of Credible Information

X X **Problems/Context:** a) People in all quarters of education display varying degrees of reluctance to use good testing & evaluation.
b) In part, a suspicion of statistics or computers exists, no matter how good the information they provide.

X X c) In some cases, people have been "burned" by the use of testing & evaluation information that actually turned out to be faulty in one or more aspects.

X X d) The research topics suggested above address the problem of faulty or technically deficient information through technical or applied perspectives while this topic emphasizes better understanding of how to gain more use of credible testing & evaluation information.

X X **Questions:** a) What barriers exist to the use of credible testing & evaluation information?

X X b) How can the discovered barriers be overcome?

X X c) How can the production of good information be made less costly?

TOPIC
T E S where T = TESTING, E = EVALUATION, S = STANDARDS.

X X **Shoulds:** a) Research in this area should help establish what the potential of the "information age" at the local school level actually is.

X X b) Research should not address general topics [illegible] what distinguishes the school as an information-seeking institution.

X X c) Any studies of information use should be tightly [illegible] to discrete issues of testing and evaluation practice.

Organization and Staffing

Specifications: Shoulds/Mays

X X 5. Staff members with practical perspectives on testing & evaluation procedures based on experience in schools are explicitly desired as part of the Center's overall staffing.

Dissemination

Specifications: Shoulds/Wills

X X 2. The Center will make its contribution to scholarly journals in testing & evaluation.

X X X 4. It will serve as a resource to those in schools & elsewhere concerned with improving schools in the areas of testing, evaluation, and standards, e.g., school district directors of research & evaluation, directors of testing, teachers & teacher organizations, state research & evaluation directors, and state assessment directors.

X X 8. Cooperation with the ERIC Clearinghouse on Tests, Measurement, and Evaluation is expected.

Collaboration with Other Centers

Context/Needs/Strengths

X X X 3. The NIE Center on Testing, Evaluation and Standards will have special skills that may be selectively used on challenges of the other Centers, e.g., it might allocate some staff time & resources to work collaboratively with another Center on a problem of mutual interest. Ideally, this would happen where both significant testing & evaluation issues and significant substantive issues were raised in a particular research problem. Other situations can also be imagined.

Examples

X X 1. Writing (writing assessment in the field)

X X 2. Reading (reading assessment in the field) not part of this grants competition

X 4. Effective Elementary Schools (methods for evaluating effective schools)

TOPIC T E S		where T = TESTING, E = EVALUATION, S = STANDARDS.
X	5.	Effective Secondary Schools (methods for evaluating effective schools)
X X	7.	Teacher Quality and Effectiveness (methods for assessing teacher competencies)

SUMMARY OF 1985 RFP FOR NIE CENTER ON STUDENT TESTING, EVALUATION, AND STANDARDS
LIST OF ONLY THOSE STATEMENTS DIRECTLY RELATED TO STANDARDS

TOPIC
T E S where T = TESTING, E = EVALUATION, S = STANDARDS.

Definitions

X **Standards:** People's values and judgments about what desirable school outcomes and processes are.

Introduction

Context

X X X 6. The perception of crisis in current educational standards is largely based upon results from testing & evaluation.

Needs/Concerns/Problems

X 2. Some of the need is a consequence of the public scrutiny of educational standards.

X 3. Although much of this emphasis is a result of federal & state initiatives, issues involving standards assume practical significance only in local efforts to improve the schools.

Scope of Center Mission

Specifications: Mission/Shoulds/Mays

X X 3. The Center ought to make demonstrable contributions toward meeting two goals: 1) a rise in test scores, and 2) a perception that educational standards have improved.

Strengths

X X X 1. Good testing & evaluation can provide information with sufficient precision, timeliness, & utility to assist in the debate & decisionmaking about standards.

Research That May be Performed

Curriculum-Test Match

X X **Strengths:** a) Further research in this area & the others mentioned above would lead to more conclusive measures of how well students are doing and what present educational standards are.

TOPIC
T E S where T = TESTING, E = EVALUATION, S = STANDARDS.

Helping with Educational Standards

X X X **Needs/Problems:** a) Current efforts to raise educational standards involve many aspects of testing & evaluation whose methodology is imperfect or undeveloped, e.g., the use of proficiency tests, etc., to regulate promotion and graduation at the local level requires instruments that are fair to all students, particularly lower-achieving ones and those near the threshold which has been established as the "passing" or "cut" point.

X X **Questions:** a) How can subject matter domains be better specified and tests made more discriminating at the critical "cut points"?

X X b) What local nonpsychometric procedures (e.g., methods for standard setting) can be developed to make pass-fail decisions more accurate and equitable?

Dissemination

Specifications: Shoulds/Wills

X X X 4. It will serve as a resource to those in schools & elsewhere concerned with improving schools in the areas of testing, evaluation, and standards, e.g., school district directors of research & evaluation, directors of testing, teachers & teacher organizations, state research & evaluation directors, and state assessment directors.

X 7. The Center's work on standards needs to be a resource to many educational and policy groups concerned with those issues at the local level.

Collaboration with Other Centers

Context/Needs/Strengths

X X X 3. The NIE Center on Testing, Evaluation and Standards will have special skills that may be selectively used on challenges of the other Centers, e.g., it might allocate some staff time & resources to work collaboratively with another Center on a problem of mutual interest. Ideally, this would happen where both significant testing & evaluation issues and significant substantive issues were raised in a particular research problem. Other situations can also be imagined.

SUMMARY OF 1985 RFP FOR NIE CENTER ON STUDENT TESTING, EVALUATION, AND STANDARDS
LIST OF ONLY THOSE STATEMENTS NOT CODED BY TESTING, EVALUATION, OR STANDARDS

TOPIC
T E S where T = TESTING, E = EVALUATION, S = STANDARDS.

Definitions

Local: Individual school districts, school buildings, & classrooms.

Scope of Center Mission

Specifications: Mission/Shoulds/Mays

2. Its resources ought to directly benefit educators and parents at the local level.

5. The primary educational interest to the Center's mission should be student educational outcomes related to achievement.

Research That May be Performed

General Specifications: Shoulds/Mays

1. The Center may conduct a mixture of basic & applied research such that the perspectives & findings of the basic & applied research will feed each other.

6. The research may involve networking of interested schools & school districts if such networking will advance the conduct or dissemination of the work. The exact mix of these components & perspectives is up to the Center.

7. In general, the Center's research should be planned to have demonstrable utility within the project period of this grant.

Information Systems

Strengths/Context: a) A small number of school districts is using the potential of technology to make organizations more productive & efficient by pursuing the development of comprehensive management or instructional information systems.
c) Different systems have been tailored for the use of teachers & other building personnel or for the use of school district policymakers.
Needs/Problems: a) The technology for efficiently gathering such information and making it accessible to all levels of a school district is not sufficiently developed or understood.

TOPIC
T E S where T = TESTING, E = EVALUATION, S = STANDARDS.

Questions: a) What is the best mixture of hardware and software for such systems?
b) How can existing technology be blended and adapted for local uses?
c) What content is most useful when included in such systems?
d) What training and other activities need to be addressed to make them workable?

Evaluation and Testing Uses: How to Gain More Use of Credible Information

Problems/Context:
b) In part, a suspicion of statistics or computers exists, no matter how good the information they provide.

Other Research

Specifications (Shoulds/Mays): a) The foregoing lines of research do not exhaust possible lines of research that may be covered.
b) All of the above line of research need not be performed.
c) Other research may be proposed if it is compatible with the mission of the Center and the topics suggested above.
d) The final selection of specific research topics should be made to maximize the Center's overall impact in its mission area.

Organization and Staffing

Specifications: Shoulds/Mays

1. No organizational arrangements are required beyond the general provisions of this announcement.

2. The Center may organize a network of schools or districts, if such networking will advance its work.

3. The Center should be organized such that the various parts contribute for each other to enhance the wholeness of its work as an institution.

4. Staff should possess expertise in the theoretical and practical perspectives involved in its research.

6. Such staff members may be employed on a rotating basis, but the overall staffing capacity in this respect must be permanent.

7. The Center should have a permanent, substantially full-time director.

TOPIC
T E S where T = TESTING, E = EVALUATION, S = STANDARDS.

Dissemination

Specifications: Shoulds/Wills

1. The following guidance is oriented toward the specific mission of this Center and it should be read along with the guidance on dissemination contained in the general provisions of this announcement.

3. It will want to be creative in getting news about its work to popular magazines & newspapers that will be read by parents and the general public.

6. The Center will be intimately acquainted with the needs of these constituencies and meet with them frequently to learn about their needs as well as pass on the results of its research.

Collaboration with Other Centers

Specifications: Shoulds

1. The Center should be alert to collaborative opportunities with other NIE Centers and plan to allocate resources to them in accordance with guidance contained in the general provisions of this grant information package.

Examples

3. Learning (higher cognitive skills)

APPENDIX 7

Summary Analysis of Concept Papers

This document provides a summary analysis of the concept and issue papers developed for consideration by the R&D Center planning team. The papers are categorized by their relevance to four R&D Center program areas. Each paper is listed by title, followed by a brief statement of its purpose. The specific issues or problems identified in the paper are noted and, in most cases, summarized. If the paper contains specific implications for proposed research bearing on the identified problem, this also is noted and summarized. In some papers, no implications or proposals for research are offered.

I. STANDARDS

A. Standards for Educational and Psychological Testing - George Madaus, Boston College

This paper deals with the Standards for Educational and Psychological Testing published by the American Psychological Association and prepared by the Committee to Develop Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. The paper describes the content of each of the four major sections of the Standards and presents implications of each section for the work of the Center.

Problems or Issues

The paper discusses each of the four major sections of the Standards. The content of each section is briefly described. It is noted that the 1985 standards will impact educational testing for some time to come, both as a result of their use in court cases and through their use by those in educational testing programs at the state and local levels.

Implications or Proposed Solutions

Part I, Technical Standards for Test Construction and Evaluation, provides the basis for the Center's development of practical techniques and checklists for school districts that are building their own criterion- or curriculum-referenced tests. Practical checklists for use by local school personnel can be developed for a variety of uses at a variety of levels.

Part II deals with the proper use of tests in different situations. The translations of the Standards into guidelines and checklists for a variety of audiences will help these different user groups appropriately interpret and use test scores. The section dealing with test use in special education is especially pertinent to the proposed Center. Also, the chapter on program evaluation may lead to other practical checklists and materials detailing the better use of test results as part of an agency's effort at program evaluation.

Part III, Standards for Particular Applications, notes that the two chapters in this section serve as the benchmark for the NIE Center. The two chapters are "Testing Linguistic Minorities" and "Testing People Who Have Handicapping Conditions." The Center should use this material to develop practical guidelines for local educators on how to modify and administer tests to students who have different handicapping conditions and who come from different linguistic communities.

Part IV is Standards for Administrative Procedures. Again, the NIE Center should use these chapters to develop practical guidelines for use by school officials regarding test administration, scoring, reporting, and protecting the rights of test takers. This paper basically suggests that the Center use this set of standards as the basis for developing checklists for materials that translate the applicable standards for users and provide guidelines for their implementation.

B. **Setting Standards - John Poggio, University of Kansas**

This paper addresses the general problem of standard setting at both the state and local levels. The paper summarizes a review of the literature that reveals that no fewer than 30 methods for setting standards on test performance have surfaced and have been used within the recent past.

Problems or Issues

1. Exploration of alternative standard-setting methodologies. Although there are many available methods, most are narrow in focus or limited in practicality. The paper suggests several problems that could be investigated in this area.

2. Establishing theories for standard setting. The paper notes that there is little theory underlying the various methods and suggests that theory needs to be formed so that we can better understand the results of the various processes or methods.

3. Implications of standard setting. A number of major and pressing problems that follow the setting of standards and deserve research attention are noted.

4. Psychometric Properties. Finally, the paper notes that research is necessary to broaden our focus to a consideration of the methods or sets of methods used in setting standards.

C. **Review of the Effects of Standardized Testing - George Madaus, Boston College**

This is a position paper, the purpose of which is a critical review of the literature on the effects of educational testing at the elementary and secondary levels. It is noted that the review is primarily concerned with the effects of traditional, school district, standardized testing programs, although the impact of newer state testing programs is included when appropriate.

Problems or Issues

The paper details several important issues related to the effects of standardized testing and highlights implications for the work of the Center. These issues are:

1. The difficulties in evaluating the literature on the consequences of testing.

2. Perspectives on the criticisms of tests.

3. Characteristics of tests and the context of testing. This section considers seven crucial dimensions along which standardized tests can be differentiated.

4. The consequences of testing. The consequences of testing are discussed under the following categories:

   a. School-level effects

   b. Teacher-level effects

   c. Student-level effects

   d. Parent-level effects

Implications or Proposed Solutions

The paper suggests one or more implications under each of the sections. The following summarizes suggested work for the Center.

1. **Difficulties in evaluating the literature on the consequences of testing:**

   a. Analyze different philosophical and ideological arguments underlying the debate about testing.

   b. Engage various stakeholder groups in dialogue to help understand the social and political dynamics of standardized testing.

2. **Criticisms of tests**

   c. Gather empirical data that clarifies, counters, or substantiates the various criticisms of tests; explicate the competing belief systems associated with support and criticisms of tests; and work with LEAs to reduce reasons for criticisms, especially related to special populations.

3. **Characteristics of tests**

   d. Work with LEAs and SEAs to help them distinguish between different objectives of measurement and to evaluate how tests can be used across different objects of measurement for different decisions.

   e. Work with school districts to develop content tests for subject matter areas.

   f. Investigate strengths and weaknesses of norm- and criterion-referenced tests relative to various uses with different audiences.

g. Help educational agencies distinguish among the various uses of test results and the various inferences they may draw from test scores.

h. Investigate different reporting strategies to communicate test results.

i. Investigate the impact of external testing programs on teaching and learning.

j. Study the levels of aggregation of test scores.

k. Be continually alert to the interrelatedness and interaction among the dimensions specified in this section.

4. **The consequences of testing:**

l. School-level effects

(1) Conduct in-depth studies in school districts to ascertain the effects of various kinds of testing programs on decision making. This study should be conducted with various kinds of districts, utilizing different testing programs.

(2) Design report forms and software packages that will provide administrators with useful analyses of test results.

m. Teacher-level effects

(1) Survey teachers from various states about their perceptions of external tests and the effects of such testing programs on the instructional program.

(2) Survey teachers working with different parts of testing programs to determine the effect of these programs on teacher perceptions of test relevance and use.

(3) Study the effects of test information on teacher expectations and teacher behavior.

n. Student-level effects

(1) Investigate the impact of providing students with test information and the effect of testing programs on dropout rate.

(2) Examine how measurement-driven instructional programs affect student behavior.

(3) Monitor the programs over time.

(4) Document effects on students of providing them and/or their teachers with test information.

(5) Develop a research program that examines test effects with different types of student groups.

- o Parent-level effects. Develop a program of research aimed at parents of different kinds of backgrounds and their understanding of different kinds of test information.

D. Professional Standards and Equity: The Role of Evaluators and Researchers - Carol Tittle

This is a paper prepared for the 1984 AERA meeting. Three major sets of standards are examined in the paper for their relevance to equity concerns in education. The standards are: 1) The Joint Committee's Standards for the Evaluations of Educational Programs, Projects and Materials; 2) The Evaluation Research Society Standards Committee's Standards for Program Evaluation; and 3) an edition of the Joint Technical Standards for Educational and Psychological Testing by the Joint Committee of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. The author also proposes equity standards for evaluators. The paper initially examines some contextual influences on professional standards and then reviews the issue of equity. This section contains many examples of how standards attempt to address equity issues. The author concludes that the test standards are responsive to equity concerns. She presents a thorough discussion of the two sets of evaluation standards with regard to their concerns over equity. The paper concludes with a proposal for a set of equity standards for evaluators.

II. PSYCHOMETRIC THEORY

A. Patterns of Item Response - M. David Miller, University of Kansas

This paper discusses recent research on patterns of item response and the kind of useful information about student test performance they can yield. The paper reviews research on item response patterns and particular indices that have been developed to measure patterns of item responses made by a student. The paper focuses especially on uses of these indices of item response patterns as a diagnostic tool. The remainder of the paper summarizes recent literature related to the indices of patterns of item responses.

B. Criterion-Referenced Testing: Research Directions - Douglas Glasnapp, University of Kansas

This paper contrasts norm-referenced and criterion-referenced tests. The paper notes that the criterion-referenced/norm-referenced distinction has become, at best, clouded. Discussion presents the advantages and distinctions of each type of test. The literature is reviewed regarding several issues involved in the norm-referenced/criterion-referenced distinction.

Problems or Issues

The point is made that there is a wealth of background information, supported by empirical data, on standardized norm-referenced measures.

Although the theory of criterion-referenced measurements indicates what characteristics test scores should have, the limited history has not provided us with comparable empirical information. Questions need to be raised and answers provided regarding what test score characteristics can be anticipated from criterion-referenced tests in the context of broader applications. Do they function as expected, based on the theoretical underpinnings of the test construction procedure?

Implications or Proposed Solutions

To answer the questions raised in this paper, research needs to be directed to specific technical issues on various aspects of criterion test construction. Research must be focused on the application of existing criterion-referenced methods. The application orientation for criterion-referenced testing research may be viewed as a focus on validity investigations into criterion-referenced test development procedures and resulting test score outcome characteristics. Therefore, research should address all phases of a testing program, including test purposes, levels of use, test development procedures, reporting formats, score interpretations, and intended audiences.

C. Test Design: Cognitive Models of Item Response - Susan Embretson, University of Kansas

The paper presents a specific proposal dealing with the design of educational tests.

Problems or Issues

The goal of the project presented in this paper is to improve the design of educational tests by explicating the cognitive processes that determine the individual's response to test items. The goal of the proposed project is further described as providing a foundation for the design of specific educational tests by assessing the cognitive components involved in important item types. A secondary goal is to study procedures in psychometric models for dynamic testing. The paper discusses the traditional practices of test design and item specifications. Two disadvantages of current practices that bear on the proposed project are presented and discussed.

Implications or Proposed Solutions

The proposed project combines methods of cognitive component analysis with latent trait models that calibrate the impact of cognitive components on item responses. The recent background research supporting this approach to studying item response is presented in some detail.

III. USE OF OF TESTS

A. The Natural Use of Student Testing and Evaluation in Schools by Classroom Teachers and Principals, and How That Use May Be Enhanced to Improve Teaching and Learning at the Local Level - James Sanders, Western Michigan University

This paper offers a proposal for studying the naturalistic uses of tests and evaluations in schools to develop prototype procedures and products for enhancing such use.

Problems and Issues

The paper concludes that neither testing nor evaluation has been integrated into the everyday practices of classroom teachers and administrators. It presents several responses to this situation and describes some naturalistic inquiry that has been pursued relative to this situation. The basic question is how testing and evaluation practices can be tailored to fit the natural information needs of school teachers and administrators to help them make better decisions and to improve the quality of instruction.

Implications or Proposed Solutions

Three procedures are offered to study the basic questions presented.

1. Case studies will be conducted in school districts of varying sizes.

2. Based on the information available from the case studies and from the literature, prototype procedures and products will be developed, aimed at enhancing testing and evaluation routines in classrooms. Each procedure or product will have student outcomes associated with it.

3. The prototype products or procedures will be prepared in final form for large-scale dissemination.

B. Issues in Testing in Bilingual Education - Robert Mendro, Dallas Independent School District

This paper addresses the problem of non-English-speaking students entering the public schools in the United States and focuses on the use of testing in bilingual education. Student testing is a major component of programs in bilingual education. It is used in student selection, the evaluation of the success of students, and as a measure of effectiveness of the curriculum.

Problems or Issues

Problems in testing students in bilingual education programs arise from two primary areas--the availability of tests and the adequacy of these tests for their intended purposes. There is a lack of tests in most other languages other than Spanish. Publishers give very little attention to this problem since the number of tests needed is so small. The problem of test adequacy, however, is the more serious problem in bilingual education. Critical concerns in this area include determining the functional level of a student's

English ability, measuring student performance against a standard in a language other than English, and adequately testing a student at his or her particular stage of development.

Implications or Proposed Solutions

Research efforts are needed in the process of learning English as a second language and related testing issues. Both research efforts and test development programs will be required to solve problems in this area and to move measurement in bilingual education to a more precise and meaningful level.

C. Use of Tests in Policy - George Madaus, Boston College

During the past five years, in a period of reform in education, tests and test results have been increasingly employed in a variety of ways, particularly at the state level. The use of tests in the policy sphere is a growing trend, according to the information presented in this paper. The paper highlights issues inherent in the policy use of tests and develops implications for the work of the NIE Center.

Problems or Issues

There are two principal uses of tests in the policy sphere. The first is the use of test information to inform policy makers, and the second is the use of tests as administrative devices in the implementation of policies. Both issues are addressed in this paper. In the case of informing policy makers, test results are used primarily to describe the present state of education or some aspect of it, or in lobbying efforts. Some of the recent educational reform reports have used test results to bring attention to what they consider the mediocre state of American education. Also, in state-wide testing or assessment programs, results are published and comparisons, whether intended or not, are made. Although decisions may be based on these tests, the paper notes that little is known about how these assessment results affect the curriculum, teaching, or learning. We do know that test results have fueled a debate on the need for educational reform and that some school districts offer programs to better prepare students to take tests.

In sharp contrast to the use of tests to inform, the administrative use of test results triggers a direct action, on either an individual or an institutional level. In many states, test results directly drive a variety of programs or actions. When tests are used in this way, the paper notes that their impact on teaching and learning at the local level is a direct function of the nature, magnitude, and immediacy of the rewards or sanctions involved. Eight positive effects attributed to the use of external tests as administrative mechanisms are presented, followed by eight negative effects. Since there are both positive and negative aspects to this use of tests, monitoring procedures need to be developed.

Implications or Proposed Solutions

The paper presents implications for the Center in both the areas of tests used to inform and tests used as administrative devices.

1. Use of tests results to inform policy

   a. Develop monitoring procedures.

   b. Study how test results indirectly impact on education at the state and local level.

   c. Identify variations in this use and study in more depth at selected sites.

   d. Based on this work, prepare materials and techniques that various groups can employ to help them interpret test data and assessment results.

   e. Use the results to plan curriculum and instructional changes.

2. Use of tests as administrative devices in policy

   f. Use the 16 positive and negative effects attributed to external testing programs as working hypotheses to guide investigations of impact at the local level.

   g. Use both survey, research, and in-depth case studies to study the effect of programs across different types of school systems.

   h. Pay particular attention to positive and negative impacts on minorities, bilingual, handicapped, and learning-disabled students.

   i. Identify and analyze uses of tests as administrative mechanisms both in this country and abroad.

   j. Develop practical materials and techniques that LEAs can use to evaluate the local effects of testing programs and improve their instructional delivery systems.

   k. Develop techniques and materials for SEAs to use to evaluate the psychometric characteristics of their tests.

D. Factors Affecting the Utility of Standardized Tests - Maria Luisa Gonzalez, Dallas Independent School District

This paper addresses a number of factors affecting the use and value of standardized testing programs in the public schools. The paper discusses purposes for testing and factors involved in proper test selection.

Problems and Issues

The paper points out five major factors to be considered in assessing the utility of tests. These are: (1) psychometric considerations, (2) administrative considerations, (3) services provided by the publishing company, (4) scoring and reporting considerations, and (5) testing costs. A number of issues underly each factor and emerge as important considerations in the utility of tests, leading to a variety of potential research problems.

Implications or Proposed Solutions

In order to facilitate the use of test results and to assure sound test selection practices, a number of potential research studies need to be undertaken. The following studies are suggested in this paper:

1. A study is needed on the degree of instructional validity and diagnostic utility of norm-referenced testing.

2. Research is needed to link testing with curriculum and instruction in a practical manner. This might involve using norm-referenced data to produce objective, reference-type data or enhancing national norm-referenced data by the addition of more specific information related to district objectives.

3. Research should be undertaken to help uncover the degree to which different commercially marketed test-taking materials actually prepare students for testing. Within this framework, the quality and use of commercial testing instruments need to be assessed.

4. An examination should be considered on how test items and responses can be made more secure nationally.

5. New avenues of testing, reporting, and scoring need to be identified.

6. Procedures that have been established in different districts should be studied to further delineate successful test program characteristics.

E. Helping Classroom Teachers Use Tests and Testing Results - Cordelia Alexander, Dallas Independent School District

This paper posits a renewed commitment to testing and, as a major aspect of this renewed commitment, recognition of the classroom as the focus of testing and test use. Teachers play a central role at this level, through administering tests, receiving the results of student performance, and using the results to guide instruction.

Problems or Issues

If testing affects students, presumably the effects often occur because teachers use test results in some consistent way. This paper addresses three particular issues that must be considered in helping classroom teachers utilize test data more effectively:

1. Issues in the various uses of this data

2. Procedures for helping teachers use test information

3. Other areas that need research

A number of issues and particular problems are addressed in each of these three areas. There is, for example, an increased responsibility to communicate the limitations of tests, to consider their effects on instruction, and

to influence proper test use. Also, teachers rely heavily on tests they themselves have developed and on their own observations of student work. Similar problems exist in helping teachers use tests properly. The paper notes seven major tasks that should be included in an assistance program for helping teachers use tests.

Implications or Proposed Solutions

In developing a series of research questions, the author of the paper offers questions presented by two other writers, Rudman and Rudner, who have summarized the following recent concerns on test use in the classroom:

- What are alternatives to testing for students diagnosis and assessment?
- How can one create a climate that fosters better test use?
- How much do parents know about testing? What are their attitudes? What are the best means of communicating student achievement?
- What are student attitudes and reactions to testing?
- Do teacher-developed tests cover a range of cognitive skill levels? Are they valid, reliable, and effective? What can be done to improve teacher-developed tests?
- Do curriculum-embedded tests cover a sufficient range of cognitive skills? Are they valid, reliable, and effective? What can be done to improve them?
- What are characteristics of useful diagnostic tests?
- How can tests be used to provide better diagnostic information about students, groups of students, and curricula?
- How can teachers develop better diagnostic instruments? How can test publishers develop better diagnostic instruments?
- What are alternative ways of reporting test results to students? Parents? The press?
- How can local school districts better analyze test results?
- How do teachers use standardized tests diagnostically?
- What are the appropriate uses for each of the different kinds of tests?
- How can the reporting of test results be improved?
- What are some successful patterns of in-service education that could be replicated for training teachers in the use of assessment data in the classroom?

F. Position Paper on Assessment for Special Education Students - Thomas J. Heiry, Dallas Independent School District

This paper focuses on the need for assessment instruments for mildly handicapped persons.

Problems or Issues

The paper, citing other sources, makes a distinction between traditional assessment, used for classification and eligibility determination, and contemporary assessment, used for individualized program planning and implementation. The implications of this distinction are especially important in the assessment of the mildly handicapped. Many issues associated with the assessment of the mildly handicapped have not been resolved and are detailed in this paper. The mildly handicapped category includes the learning disabled, the educable mentally retarded, and the slow learner. Many problems exist with the assessment of these students, since they are often assessed in conjunction with the more severely handicapped. The utility of traditional psychoeducational assessment has been questioned on grounds that (a) etiology cannot be reliably determined using traditional measures, (b) etiology has been shown to be linked with a type of educational intervention, and (c) traditional assessment for special education eligibility is costly.

An assessment of the mildly handicapped is best viewed as a two-function process. Traditional psychoeducational assessment is utilized in conjuntion with other data about the students to make critical decisions concerning the presence or absence of a handicapping condition. Contemporary curriculum-based assessment is utilized to make individualized, educational, and programming decisions.

Implications or Proposed Solutions

The research questions are viewed in relation to an understanding of special education assessment as a two-function process. The research questions address diagnosis and how diagnosis leads to better contemporary assessment.

1. The major task for research in the field of special education assessment is the operationalization of what constitutes a reliable and valid diagnosis for learning disabilities.

   A. What are the variables that discriminate a learning-disabled student from other problem, or handicapped, learners?

   B. What is a clinically useful approach to determine a discrepancy between ability and achievement?

2. The design and psychometric properties of curriculum-based tests in this area is another area for research.

3. "Exit criteria" for dismissing a student from special education after instructional efforts have been implemented is another area for future research. Correlates of successful mainstreaming should also be examined.

G. Implications of Special Needs Populations and Assessment Practices - Edward Meyen, University of Kansas

This is another paper addressing issues related to problems of handicapped children and youth. The focus of concern is the appropriateness of assessment instruments used, and the determination of eligibility for placement, in special education programs. The paper clearly and cogently discusses the special needs population and the academic context. This discussion provides a frame of reference for exploring assessment-related research issues as they pertain to the special needs student population.

Implications or Proposed Solutions

The paper provides the following list of statements that can be translated into research questions to provide the basis for systematic investigations of how best to assess the academic performance of special needs students.

1. Children with disabilities tend to be subjected to more testing than other students, but, except for the mildly handicapped, they are rarely tested in basic academic skills.

2. Because of the variability within school districts in policies governing the administration of academic skill assessment tests as they apply to students with disabilities, aggregate results of testing programs at the district level may not be representative of students enrolled.

3. Teachers of mildly handicapped students are oriented toward teaching for mastery and thus tend to favor a criterion-referenced testing approach.

4. The assumption is often made that oral administration of tests to this group compensates for the limitations they may have in reading. Results of research in this area are mixed.

5. For the mildly mentally retarded, and to a certain extent for students who are emotionally disturbed or learning disabled, lack of motivation is considered to be a major contributor to poor academic performance. A student's history of failure in learning and testing situations causes the student to accept failure and not to comprehend the significance of assessment exercises.

6. Teachers often exercise care in the amount of work assigned to a student for completing in one sitting and may also provide prompting in this work in the instructional program. A modular approach to the presentation of test items and allowance for some level of prompting merit exploration as methods of building confidence in testing situations.

7. There is a tendency with handicapped students to introduce the application of academic skills into the curriculum as early as possible. The concern for application is commendable, but it may be occurring at the expense of teaching the necessary basic skills and may limit the scope of concepts taught.

8. The language structure used by children with intellectual impairments is less developed than in non-disabled peers. In developing test items, reading difficulty often requires primary attention. Simplification of the language structure when assessing more complex skills may be effective with this population.

9. Computerized testing procedures that combine testing with remediation should be explored.

H. Legislative Impact and Educational Reform - Florida PDK Consortium

This paper is a series of related documents that resulted from the work of the Florida PDK Consortium Planning Task Force. it reviews the history of programs in Florida, among which are various testing programs for students at all levels. It summarizes the concern of a number of education and societal groups regarding the programs and references, criticisms and issues, that have appeared in the public media. It presents a plan for collaborative study of the programs and of the process by which such programs can be implemented in Florida. The paper should be read in its entirety for full understanding.

The State of Florida has been at the forefront of legislating educational reforms during the past decade. Beginning with the Educational Reform Act of 1976 and continuing through the Management Training Act of 1983, legislative decisions have focused on the setting of student standards, teacher effectiveness, administrator effectiveness, and curriculum and instruction.

As Daniel Duke has noted in his recent essay on educational excellence,

> Hard on the heels of most reform movements in education come the demands to know whether or not the reforms have "made a difference" (Duke, 1985, p. 671).

Accountability for a major investment of resources into legislated reforms is certainly one reason for undertaking an evaluation of recently legislated educational reforms in Florida. An even more compelling rationale for this study, however, is the need to learn from past experience so that current and future efforts at reform may proceed from an expanded knowledge base.

There are reform packages already near completion in several state legislatures, and more are planned for the coming year (Pipho, 1985). Current legislative and state board activity in bellwether states is proceeding without benefit of systematic and thorough understanding of the consequences of past efforts for teaching and learning. Since student testing and standard setting have been used as instruments to bring about desired changes in Florida, an evaluation of these efforts will have significance for the policy-shaping communities of every state considering the use of testing and standard setting to reform its educational system.

Thus, the research problem to be addressed is, "How have legislative testing and standard-setting programs in Florida affected significant stakeholders and critical institutional functions in the education system?"

The purpose of a project would be to collect empirical information about the impact of student testing and standard-setting mandates or acts legislated

in one state, Florida, that has been active in education reform, and the impact of its testing programs on important stakeholders in the state and on institutional functions within the state education's system.

Specifically, this project could result in:

1. descriptions of documented effects found in selected schools and school districts

2. descriptions of changes in student needs and achievement over a ten-year period

3. recorded testimony and analysis of this testimony for individuals and groups affected by the legislation or mandates;

4. analysis of reports collected within Florida and from other states in which the impact of mandated or legislated testing and standard-setting programs have been studied

5. survey of findings of reported impacts on stakeholders and institutional functions throughout the state;

6. interpretations and recommendations by an expert panel on the use of testing and standard-setting legislation and state mandates to improve the education of students

7. a model for state-wide evaluation of educational reforms through mandated or legislated testing and standard setting

Such a project could provide answers to the following questions:

1. What documented impact was found for testing and standard setting legislative acts and state mandates on students, parents, administrators, communities, and business/industry?

2. What documented impact was found for testing and standard-setting legislative acts and state mandates on educational policy making and planning, finances and facilities, curriculum and instruction, and school/community relatins?

3. What aspects of state mandates and legislation appear to have the greatest impact on stakeholder groups and institutional functions?

4. What research-based recommendations can be made to states regarding current and future legislation and mandates that involve student testing and standard setting?

5. How can states best evaluate the impacts of legislation and mandates that involve student testing and standard setting?

I. Influence of Training in Measurement Skills in Higher Education by Mary Anne Bunda Western Michigan University

The paper presents the background and structure for a research project dealing with the influence in training measurement on the uses of test in school. Background is presented for the need to improve uses of test based on current research, writing, and state activities. The paper presents four propcsition related to testing and assessment in schools. Two of the proposition are relevant generalizable to post secondary institution. It further points out that higher education in general is not prepared to deal with training the area of evaluation and assessment. Evaluation work needs to be undertaken both the current status and the use of tests in higher education.

Implications For Projects

There are four objectives for the research proposed in this paper. (1) a preliminary identification of patterns of test constructs and use will be investigated in higher education. (2) the identification of barriers to training in testing in higher education will be attempted. (3) an analysis of programs designed to promote testing in higher education will be completed (4) Models for training in higher education based upon the finding from this project will be developed.

The project is conceived as a one year project using naturalistic inquir techniques. The study will result in reports on research activities with can be disseminated through the center, structured instrumentation which can be used to study the same phenomena in a larger sample of institutions, and training material that can be used by center personal or disseminated for broader use in higher education facilities.

J. Judicial and Legislative Influences in Educational Testing - Diane Pullin

This paper is concerned with the influence courts and legislatures have upon educational policy makers, especially in light of the increased uses of tests for classifying students. The paper first discusses the involvement of the courts and legislatures in educational policy and describes relevant cases. Courts have begun recently to scrutinize the tests themselves and this development also is reviewed. Most recent court cases have resulted from testing requirements that have been legislatively initiated. Two primary areas addressed by legislation have been tests used to determine program or individual educational accountability and tests used to determine eligibility for special education services. Cases and issues relevant to this latter are described in some detail. The paper then identifies nine legal issues emerging from legislative and judicial decisions.

K. Equity Issues and Student Testing - Arnold Gallegos, Western Michigan University

This paper addresses one important problem associated with increased use of testing to assess student progress and to classify students as part of the effort to correct problems with instructional programs in the schools.

Problems or Issues

The primary focus of this paper is on a series of equity questions. The first issue described in the paper is curricular equity--the balance in teaching the basic skills which are tested and cutting back on other areas of study. The issue of what subject areas received top priority also leads to problems of equity in professional roles. The third area discussed is that of social and cultural equity and the impact of testing on this issue.

Implications or Proposed Solutions

Two research issues of relevance to the proposed center are suggested; (a) work on test, construction and response modes; and (b) assessing the impact tests have on the curriculum and teachers.

IV. THE DEVELOPMENT OF NEW SYSTEMS AND TECHNOLOGIES

A. Computer Assisted Professional: A Proposal to Develop an Information System to Assist School Professionals in Planning and Implementing School Improvement - William Cooley, University of Pittsburgh

During the 1984-85 school year, members of the Learning Research and Development Center have been working with the Pittsburgh Schools in the development of a prototype microcomputer system in one elementary school. This paper presents an analysis of the need for an automated information system in a local school. It traces the use of automated information systems from the 1950s and 1960s, and discusses applications for centralized, system-wide functions. It is proposed that, with the arrival of inexpensive but powerful microcomputers, automated information systems can be used effectively in local schools. The general goals of such systems are (1) to improve student achievement, (2) to enhance the quality of school life, and (3) to provide equal learning opportunity for all students. The functions of the automated information system at the local school building are also discussed.

Problems or Issues

To make an information system useful at the local level, its characteristics need to be designed specifically for local use. Issues of dependability, accuracy, and currency of information files are discussed. The integration of several information functions at the local level is presented as an important outcome of the developmental work. With these problems in mind, staff of the LRDC designed and implemented a prototype system in one elementary school during the past year. The feasibility of such a school-based information system for improving information flow and information use has been demonstrated; however, considerable work needs to be done if such systems are to operate without the day-to-day presence of a technical support team. Also, the job of building into the system the kind of expertise that will assist the professionals in each school still remains.

Implications or Proposed Solutions

The proposal presented in this paper is to further develop and test this prototype system in the context of the Pittsburgh Schools. Two major tasks are proposed:

1. to carefully test the notion that the system currently in place can be operated successfully by personnel in the school with no additional staff requirements; to place the system entirely in the hands of the school personnel and to study what happens

2. to replicate the initial implementation in another school in the Pittsburgh Elementary School System

B. Computer-Generated, Personalized Testing - George H. Olson, Dallas Independent School District

This paper briefly presents the possibilities of developing and using school system computer data bases for student evaluation. The writer notes that the ideas presented can only be developed in a context of a large-scale research and development effort.

Problems or Issues

In large school districts, the current use of centralized information systems and the linking of student evaluation to test scores on nationally and locally developed standardized tests presents special problems in the adequate evaluation of students. Student evaluation generally is reduced to an annual set of observations. Further, the validity of these observations with respect to individual students depends to a great extent on the correspondence among the district's curriculum and instruction objectives, the student's level of ability, and the skill and ability domains measured by the test. This often results in a less-than-adequate assessment of individual students and a needless waste of valuable instructional time.

Implications or Proposed Solutions

The paper proposes an efficient system for generating unique, personalized tests tailored to individual student characteristics, but which still retain the properties of standardized tests allowing for normative or comparative evaluation. In the system proposed, the construction and printing of individually unique tests would be accomplished on the mainframe computer. The system would be based on recent technological advances in test-related fields such as item response theory, adaptive testing, computerized item banking, matrix sampling, etc. The paper presents a brief overview of the system and discusses the advantages of the approach.

C. The Program on Technology in Student Testing, Evaluation, and Standards - Richard Frisbie and Lyke Thompson, Western Michigan University

This paper presents opportunities that currently exist for applying technological applications to testing, evaluation, and standards in local settings.

Implications or Proposed Solutions

Three potential projects are suggested in this paper:

1. Research and development on school district data bases. The specifics of this project are presented elsewhere in the institutional proposal.

2. The expert advisor series. This proposal suggests an "expert systems" approach for developing technical assistance materials aiding consumers at the local level. The expert adviser series of materials would be a growing collection of expert system computer programs dedicated to helping producers and consumers at the local level better understand, use, develop, and improve testing, evaluation, and standards for student academic performance. The activities suggested for this project would include (a) the identification of areas of expertise in testing, evaluation, and standards that can be provided to local users and that can also be implemented on a microcomputer-based expert system, (b) the selection of at least two areas for prototype development, (c) the development of the expert adviser in these two areas, (d) field tests of the expert advisor materials and subsequent revision, (e) dissemination, and (f) development of new titles for the series. Two examples are provided in detail for the reader.

3. A programmers' library for computer-based testing and evaluation. This project consists of the development of a specialized microcomputer program library on testing and evaluation for use by consumers in the field. The project would focus on the necessary criteria and judgments for the selection of the materials. Standards for use in this project would be developed.