EDUCATIONAL TESTING AND EVALUATION: CRITICAL RESEARCH AND DEVELOPMENT NEEDS

Joan Herman
Editor

CSE Report No. 213
1983

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Table of Contents

Introduction

Joan Herman

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

The UCLA Center for the Study of Evaluation (CSE), as part of its mission as a national center, seeks to identify significant problems in the field and to illuminate a national agenda for research and development in educational testing and evaluation. Committed to a broad-based and collaboratively developed research agenda, CSE sponsors a number of activities to provide a forum for a wide range of interested constituencies. This document reports the result of one such forum, a testimony session conducted at the annual meeting of the American Educational Research Association in April, 1983.

Representatives from a broad base of national, state, and local education agencies were invited to express their views on critical needs for research and the role of educational testing and evaluation in solving pressing educational problems. Representing national associations of educators, state departments and boards of education, local school districts, and colleges and universities, a number of common themes emerged from the testimony. Each of these themes is briefly summarized below.

Evaluation systems, testing, and technology emerged as three commonly identified research priorities. Speakers noted the need for research on the development of comprehensive evaluation systems for local schools. In their views, such a system should operate at the school building level and contain a broad base of information on

school and classroom practices, a wide range of outcomes, student characteristics, and contextual factors for use in better assessing student and school needs, providing appropriate instruction, effectively managing the school site, and promoting educational quality and accountability.

Testing priorities concerned the search for alternatives to standard measures and strategies to broaden the focus of assessment as well as technical measurement issues. Speakers noted the need to develop multiple criteria and procedures for assessing the performance of students and schools including both quantitative and qualitative measures responsive to individual differences among students. Concerns for equity were widely emphasized. Speakers also called for the development of sound classroom relevant measures in domains beyond basic skills such as higher-order thinking skills, and in secondary curricular areas such as algebra and French. Technical issues such as measuring change, sampling domains, and vertical equating were also identified as important areas of inquiry.

Inquiry into the application of technology, particularly microcomputers, was also a dominant theme in the testimony. The need for research-based guidelines for selecting and using educational software, both for instructional and management purposes and for the development of appropriate computer literacy curricula, was consistently mentioned by speakers. Many speakers recommended tapping the power of computers to administer and score tests, to provide immediate feedback on performances, and to decentralize capabilities to local schools.

Research in teaching and learning, planning and management, and inservice/preservice training also emerged as research priorities in the testimony. Recommended priorities in the area of teaching and learning included inquiry into the basic principles of teaching and learning, the applications of behavioral and social science findings to education, and the identification of techniques and practices that enhance learning.

Speakers suggested research on the development and application of techniques to serve the immediate decision-making needs of local schools, e.g., projections and simulations of enrollments and other demographic factors for use in making decisions about budgeting, desegregation, scheduling, school attendance areas, and school openings and closings. Management studies were urged to focus attention on strategies for effectively managing finances and programs at the district, school, and class levels. Research on strategies for effective training, both preservice and inservice, also emerged as an area of concern. Research-based training programs for administrators, teachers, and members of research/evaluation units were identified as focal.

In addition to the substantive priorities noted above, testimony focused on the need for cost-effective research strategies that provide generalizable as well as site-specific information. One representative suggested a "standard parts" and "variable parts" strategy that would provide standard research specifications across sites as well as site-specific variations based on unique local features and considerations.

Finally, there was general agreement that advances must be made in disseminating information from research studies to practitioners and in fostering collaborations. Decision makers must be better informed about research findings and their applicability, and better collaborative relationships need to be developed among institutions of higher education, R & D agencies, and state and local school systems.

Improved and Expanded Assessment Techniques

Sharon Robinson

National Education Association

The NEA has a longstanding committment to excellence in all dimensions of education, including testing and evaluation. NEA's priorities, programs, and activities over 10 years have included those to bring about more rigorous, accurate evaluation of student learning progress.

NEA supports, in principle, the identification of "a national agenda for research in educational testing and evaluation." We support this concept in principle because:

1. The need for solutions to the critical problems of evaluating student learning progress is great and far transcends state and local conditions.
2. State and local jurisdictions (state education departments and local districts) do not have the resources, capacity, or will to conduct such research.

Pressing Problems in Education

1. Equity--current testing practices treat students of different backgrounds differently.
2. Opportunity--present evaluation and testing practices narrow rather than broaden student opportunity.
3. Accuracy--present evaluation and testing practices either measure school purposes too narrowly, or distort those which they do measure.

4. Instructional strategies--current evaluation and testing practices foster questionable pedagogical practice, e.g., teaching to the tests.
5. Public understanding--present evaluation and testing practices result in the public (and policymakers) being misinformed about the quality of schooling.

Areas of Priority for Research

1. What value (weighting) should be placed on assessing the various purposes of education: preparation for citizenship, preparation for work, and personal development? There are others.
2. What emphasis should be placed on evaluating the person in the process as compared to assessing outcomes only?
3. What criteria for assessing student learning beyond paper-and-pencil tests hold the most promise for future development?
4. How can those "promising others" best be developed and used?
5. How can multiple criteria be used so that no single criterion becomes the "go/no-go" determiner in making significant decisions about students?
6. How can qualitative evaluations such as case studies, observations, and interviews be interpreted and reported so that they are as favorably received and productively used as paper-and-pencil assessment?
7. How can practitioners be trained to use a broad range of evaluation processes and instruments?
8. How can policymakers, administrators, and other decision-makers be educated to accept and value a broad range of evaluation criteria and to demand that multiple criteria be used?

9. How best can evaluations be accomplished that do not employ grade equivalents?
10. How can evaluations be accomplished that take into account differences in learning rates?

Some Caveats in Moving from Principles to Programs

1. If the posture remains that assessing students is like measuring hard goods products--that improving quantified measurement is the important objective--it won't be worthwhile.
2. If the approaches are to be mainly improving those processes and practices outlined in Standards for Educational and Psychological Tests (now under revision), it won't be worthwhile.
3. If the emphasis is to be on research to improve current state assessment programs, it won't be worthwhile.
4. If the emphasis is to be on "tinkering" with any form of standardized measurement, it won't be worthwhile.

In Conclusion

What is needed is a new order in evaluation and testing:

1. One that begins with a concern for increasing rather than reducing opportunity.
2. One that moves away from psychometrics to more qualitative concerns.
3. One that gives much, much more attention to the person in the process.
4. One that gives increased consideration to individual students' styles and tempos of learning.
5. One that judges a student's progress against objectives to be accomplished rather than against other students.

Research-Based Decision Making

Larry Barber

Phi Delta Kappa

Education, quite different from business, industry, and many other forces of human inquiry, does not utilize a research base for decision makers. Most educators have never been taught to use data-based decision-making systems. At the same time, education has never developed a research capability or a research utilization attitude.

On the other hand, educators consider the political decision-making process to be paramount. This attitude must change. Funds for education research must be expanded. Funding a research-based system that will provide a valuable adjunct to the political decision process will allow educators to see an exemplary combined decision-making system work.

There are three major areas where funds for educational research must be directed: basic research, field research, and a search for practical applications of research.

Basic Research

Funds for research must be directed toward examining the basic structure of education and its relationship to learning theory and to the body of knowledge in the social sciences. Basic research in education should hold a prominent place on the national agenda because, through such research, practice is eventually improved.

In education we often do not know what we should do. Basic research is what you should do when you don't know what you should

do. Some of our research funding must be apportioned to discover what we should do.

a. Research that attempts to identify the basic principles of learning and teaching.

b. Research that tests the application of behavioral and social science findings to education.

c. Research that identifies the basic techniques and practices that enhance learning.

Field Research

Research funds must also be directed toward identifying and disseminating exemplary practices in education. Such practices may involve teaching, managing, developing curriculum, or implementing new programs.

Specific Agendas for Field Research

a. Research that identifies the basic properties of successfully managed school districts.

b. Research that identifies the basic properties of successfully managed classrooms.

c. Research that describes what actually happens in a normal, good classroom.

d. Research that identifies and describes the behaviors of an outstanding principal.

e. Research that identifies the properties of an excellent teacher-training program.

As a part of our efforts to identify and disseminate exemplary practice in education, we at Phi Delta Kappa have polled our

membership and the members of other education agencies to determine areas of education where they feel research data would help them solve the problems they face. The results of the polls indicate needed field research in the following areas:

1. declining enrollment and related problems;
2. differentiated pay or merit-pay systems;
3. principal effectiveness;
4. optimal ways of using microcomputers in instruction;
5. optimal fiscal management of school districts;
6. optimizing the in-service training program;
7. reduction of student absenteeism and dropouts;
8. identification of realistic student competencies.

We suggest that these problems and areas of interest warrant research funding.

Practical Applications of Research

A major area of neglect that could rightfully be called research, but more appropriately might be called engineering, is the concept of practical implications of research in education. We do not, however, have a system that allows educators to learn basic research applications or field research, nor do we have a system to explain to researchers the information that practitioners need to improve practice. We recommend the following remedies.

Specific Agendas for Practical Applications of Research

a. Research in identifying the best ways to encourage educators and researchers to communicate common issues and problems.

b. Research in identifying better ways to train educational administrators at all levels to use research information.

Linking Policy Makers' Needs with Evaluators' Skills to Promote the Wise Use of Test and Other Evaluation Information

Gerald W. Bracey
Director of Research, Evaluation and Testing
Virginia Department of Education

There are so many areas in the fields of testing and evaluation in need of research that I feel practically paralyzed deciding where to start. I will limit my agenda to activities derivable from the following six somewhat related points.

1. Standardized test scores (usually percentile ranks, not actual scores of any sort) are becoming more and more visible in the media, usually in terms of declines or "box scores" reporting that X school district did Y% better (with percentile ranks!) than Z school district. In many cases, such scores or ranks represent the sole credible piece of information for evaluating the quality of instruction in schools.
2. The increased use of test scores has not yet been accompanied by an increased understanding of their appropriate and inappropriate uses, or their meaningful and meaningless interpretations.
3. Other and more appropriate methods of evaluating instruction than norm-referenced test scores are being overlooked.
4. There is a great deal of discussion about the need, in an information society, for citizens who can think and reason well.

5. The research on the development of competence and expertise is developing exponentially but is still in a relatively primitive state.
6. Test development has been based on factor analytic techniques developed many years earlier and not examined for current relevance, or has been based, as in the case of achievement tests, on simple content analysis of existing curricular materials. Little test development has been based on theory or data derived from developmental or cognitive psychology.

There are other points that could be made, but these should be sufficient for a rather ambitious research and development agenda. Part of the agenda is "basic" research, and part is "applied." As noted, some of the points are related and will be considered together.

Given the increased use of test scores for a variety of purposes, both appropriate and inappropriate, it follows that there is an R & D agenda to insure that policy makers--who are seldom those who construct or administer tests--are better informed about the uses and limits of the various types of tests. Ditto for the media. At the very least, a series of short, readable papers such as Donald Horst's "What's Bad About Grade Equivalent Scores" should be prepared, covering other statistical properties of tests and issues concerning the interpretation of test scores. Horst's paper is ten pages long, which is probably the maximum for the audiences being considered. Indeed, an aural or audiovisual presentation would be preferable. Many people in policy-making positions are aliterate: they haven't the time to read anything extensive and have developed the trait of

learning from oral briefings and making decisions based on such briefings.

The task will in no case be easy. At many newspapers the education beat is for cub reporters, and few radio or TV stations have an education reporter. The copy of reporters is often altered by editors and totally distorted by headline writers. Moreover, many persons in a position to inform legislators about testing and evaluation are constrained by protocol from doing so (they are also constrained by the limits of their perspectives, which will be considered below). The fact that it is not an easy task doesn't mean that it is not one to undertake.

Concomitant with an attempt to better inform people in general about tests should be an attempt to wean them away from tests as the sole criterion for evaluating schools. The effective schools literature, for all its many flaws, is a step in this direction, but only a step. Research should be undertaken to develop instruments for measuring other attributes of schools as "objectively" as achievement is measured by tests. An analysis of why test scores are credible should allow the development of a set of criteria for other indices and other evaluation techniques.

An attempt to disenchant policy makers and the media (and hence the public) with tests as the only evaluation technique should be accompanied by research on how researchers can reach policy makers. Many researchers-evaluators-psychometricians are as unaccustomed to giving succinct briefings as the policy makers are accustomed to receiving them. They are used to presenting lots of data which should

"speak for itself," whereas policy makers are looking for conclusions already drawn. Lee Cronbach addressed this issue to a limited degree in his remarks to the initial meeting of the Evaluation Research Society, but many, perhaps most, evaluators see themselves as doing scientific research for which there is no timeline except, perhaps, the duration of the grant. Policy makers live in a world where decision points far outstrip the researcher-evaluator's ability to provide information that is acceptable to the researcher in terms of scientific rigor and reliability. Researchers and evaluators need to ask not "How long will it be before I can have a definitive answer?" but "What kind of answer can I provide by a given deadline?" Research into techniques for this "real time" research and evaluation should prove interesting and useful, and it seems vital.

One possibility for speeding up the process is through the use of microprocessor technology. I am unaware of research or development efforts using computers to speed up the flow of information, aside from the use of word processors to write the report when it's already too late. Surely there should be some utility in the existing data-based management programs, and if these prove to be of limited value, just as surely there would be a market for the development of such systems for rearranging test and other evaluation information. Such rearrangements, quickly and easily done, are necessary because it is practically a law that whatever data exists will not be in exactly the form that the policy maker wants.

I want to shift my conceptual reference point at this time and deal more with higher order thinking skills. After a few years of

being overwhelmed by minimum competency, we are now similarly overwhelmed by calls for "excellence" and citizens who can think and solve problems in an information society. When we look at the literature for tests in this area we find that the cupboard is not exactly bare, but it isn't well-stocked with nutritious items, either. With psychology split from education for a number of years, with experimental psychology chasing a defunct model of physics in search of respectability, and with educational research chasing experimental psychology in search of respectability by association, a literature of the assessment of many skills has grown up uninformed by developmental psychology or the more emergent cognitive psychology.

The work of people such as Sternberg, Glaser, and others doing research which is generally referred to as componential analysis offers some hope of changing this situation, but the data base is as yet very small. Much more research is required on how competence and expertise is developed, and on how to assess it. There are many promising starts, ranging from artificial intelligence to "brain-based curricula" to the rediscovery of Lev Vygotsky's "zone of proximal development," but they are no more than starts. In some instances, particularly those derived from research on brain functioning, research needs to be conducted to verify (and probably disprove) many of the conjectures being made about the assessment of right brain/left brain functions.

Finally, there needs to be much more research on the kinds of tests we really need. There is work in progress at several institutions on adaptive testing, but this work to date has been

limited to multiple choice items and has required a minicomputer. Research on such test development procedures should be continued, but should be extended to obtain more information from the student than optimal level of difficulty. For example, it might prove worthwhile to use a computer with real-time clocks in the program to determine how long a student ponders over a problem or parts of the problem. In a somewhat different vein, research is needed on existing norm-referenced tests which purport to also provide criterion-referenced information. I and my colleagues at the Department of Education in Virginia are concerned that the attempts to weld these two functions together have muddied conceptual waters, and that the tests that purport to do both may, in fact, do neither well.

Over three years ago I proposed an article to Jim Popham, then editor of Educational Evaluation and Policy Analysis, based on 13 factors which I saw as reducing the relationship between evaluation and policy. One of the factors listed was entitled "Alligators and Brushfires: The Unexamined Life in a Bureaucracy." It referred to the fact that in many bureaucracies the pace of activity is such that one does not have time to stop and ask the original purpose of the activity. Jim responded positively to the ideas but, unfortunately, I fell victim to the brushfire syndrome and never wrote the article. On reexamining the list, however, I find many of the factors still operating. Some of them have been touched on already, and a number of them provide seeds for a research agenda. I repeat them here for your amusement and, perhaps, information.

Some Factors Which Reduce the Relationship Between Evaluation and Policy

1. Stop the world, I want to conduct my evaluation: fixed variables in a world of constant flux.
2. The hills are alive with the sound of trial balloons: statements made for the purpose of seeing who supports and who opposes whom are often taken too literally by evaluators.
3. White hats, black hats, and hired guns of various stripes: role confusion and divided loyalties in evaluators.
4. Concrete operations and brick walls: evaluators' egocentrism often leads them to examine wrong variables with inappropriate methodology.
5. The view from the Ivory Tower, or, Fear and Loathing in the Poison Ivy League: the disdain of many evaluators connected with universities for the people in SEAs or LEAs leads to mutual distrust.
6. The total morality of the straight line, or, linear thinking on the road to nowhere: attempts to fit round problems into square methodologies lead to irrelevance.
7. Raising the roof and razing the organization: because whole organizations will not disappear, evaluators sometimes despair of doing any renovation.
8. Rendering sterility from chaos--how many angels can boogie on a grade equivalent: the necessity for simple, quantifiable outcomes overlooks richness and the complexity of the problem.
9. Newton's first law of motion and Old Man River: the power of inertia. How much impact is enough?

10. Alligators and brushfires: the unexamined life in a bureaucracy.
11. Professional, personal and institutional loyalties: can a house divided stand up for itself?
12. Public hearings and policy: the spoils of legislation go to the most vocal.
13. How do you spell relief? P-O-W-E-R: evaluation innocents abroad in a world of competing power structures.

Depth and Breadth in Research

Norman Stenzel

Illinois State Board of Education

Indeed these may be the best and worst of times. Critics of education have pointed to issues which could stimulate significant research and evaluation efforts. These include the quality of education, the federal role in education, and micro-computer technology in education. Schools face conditions which rend the best run systems to their very foundations. Such conditions include budget cutting marathons, reductions in force, debate over the content of curricular offerings, battles over the nature of graduation requirements, and confrontations with pressure groups. While these are terrible times for schools, these would be the best of times to examine the environment of education as an experimental treatment. Pragmatic studies would inform decision makers about these conditions. Unfortunately, researchers and evaluators even now are facing renewed criticism and attacks upon the utility of social science. Response to either the needs or the attacks is complicated by the allocation of fewer funds to implement work.

Areas of Priority for Research

One response to these difficult times could be to focus on a few issues, spending a significant amount of available monies (a fiscal "critical mass") to assure that the best possible study is implemented. Another response might include funding a larger number of small studies to cover as many issues as possible. Neither approach is entirely satisfactory. Although the reduction of the

number of issues to be examined could be done through prioritization processes, priorities are rarely identified accurately when compared to hindsight. On the other hand, the small study alternative often runs aground because of constricted scope or limited fiscal resources.

There are a number of strategies which may be useful during these difficult times. The strategies can cope with limited funding, sampling issues, and methodological concerns. The labels used here are a "standard parts" and a "variable parts" strategy.

Standard Parts Strategy

The large-scale study often overlooks the individual component result. To obtain cooperation from local educational agencies, direct payoff must be incorporated into large-scale efforts. A strategy which designs a study to be implemented at the individual district or school level according to standard specifications (allowing approved variances) would provide local results plus multiple replications for comparisons to other studies conducted in the same manner. Analyses could be conducted at the local level if resources are available or at a headquarters location when deemed appropriate. Individual study results could be reviewed wth the best current meta-analysis techniques to provide the large-scale effort result.

Meta-analysis in this case could be taken in a much more general sense than the Glass, McGraw and Smith (Meta-Analysis in Social Research, Beverly Hills: Sage, 1981) usage. While they attend to quantitative studies, the intent here is also to allow processing of qualitative data such as in the manner suggested by Yin and Heald in "Evaluating Policy Studies by Using the Case Survey Methods" (Santa Monica: Rand, 1975).

The standard parts design is not suggested as a substitute for other large-scale study strategies. Indeed, it may not be widely used where topics are not amenable to local level results.

Variable Parts Strategy

While a standard parts approach seeks to obtain multiple implementations of the same design as a way of providing depth of data, it sacrifices breadth. The variable parts strategy would seek to obtain greater breadth.

The variable parts approach would build in important variations in the treatment condition to be analyzed in a quantitative or qualitative manner. If qualitative study results are used, the analyses could seek to identify common or unique features of each case. A profile of successful conditions, as opposed to unsuccessful or less successful conditions, could be a feature of the general report. The local report could summarize the presence or absence of the study-identified conditions at the individual site.

Implementation

Quantitative or qualitative meta-analysis of multiple standard or variegated studies will have some common features. While critics may decry loss of control and the problems of obtaining samples through voluntarism, those issues can be met.

The director of these projects would be like a composer and an orchestrator. The research design, with all of its specifications and conditions, would be the musical score. Allowing for local circumstances would be the task akin to that of the orchestrator.

The research undertaken would have to be of great interest to local personnel. Only studies with local payoff are likely to get off the ground. Specific reports of results for local use will have to be built into the work undertaken. The funding for these strategies could also emphasize the decentralized character of the work. Local funds might be possible in districts where local resources have not yet dried up. On the other hand, start-up monies for such projects as computer-enhanced education programs could also include evaluative funds which could be applied as a part of either the standard or variable strategies, thereby multiplying local information.

At the End

While we all might want to gain glory through a major study triumph, the odds are against it. Critics abound. Even the strategies briefly sketched here are too vulnerable to critics. However, the importance of the utility of the parts of the overall study will bring favorable support from parties previously neglected. It is worth a try.

Identifying Decision Areas That Make a Difference

Steven Frankel

Montgomery County, Maryland Public Schools

I think a major problem with the research agenda in educational testing and evaluation is its narrow focus. We've been trying to mine a very, very limited area for a long period of time; it's time to broaden our horizons to other areas that can really make a difference in school district decision-making.

Demographic Analyses

One example is the need for demographic analyses. We need to develop better techniques for identifying where students do and will reside, and for simulating what might happen when students are shifted around. These sorts of analyses would provide information that is needed right now for budgeting decisions, desegregation decisions, and decisions about bus schedules and school closings. Yet I'll wager there isn't a person here who has any training or background in demographics. This is an area where we need to develop better techniques and where we need to provide training.

Management Studies and Audits

Another area where we likewise need to give more attention and training (and one I've been pushing for a few years) is management studies and audits. Forty percent of every education dollar goes into non-teaching areas, yet we virtually ignore these areas in our research and evaluation efforts. Every nickel that we can save from this "other" forty percent could be put in the classroom, yielding direct and phenomenal political pay-offs.

Not only is there a potential high pay-off, but it makes good sense for the research unit to take over the responsibility for management studies and audits. There are serious conflict of interest problems when the director of accounting conducts audits of a school system's books and its management. But if the R & E unit is going to be involved, then we need to develop the appropriate techniques and find people who know management and accounting or provide appropriate training.

We need to widen our role as evaluators. We need to go into a program and look at its books to see how money is being spent and whether the operation is being managed efficiently, as well as investigating whether educational outcomes have been reached. We've put ourselves in a box that is so narrow that we're essentially irrelevant to most educational decision-making. We need to get out and widen our agenda.

Comprehensive Assessment Systems

We also need to do long-range work on school-building-level assessment systems--systems that go way beyond testing. Testing, of course, will continue to be a part, but we need to broaden our purview to look at community satisfaction, school visitation, school practices and climate, and build comprehensive systems for looking at what's going on at the classroom and building levels.

Follow-Up Studies

A fourth area is the need for follow-up studies to assess the quality of schools. It is impossibly expensive for a given LEA to reliably follow-up graduates five and ten years out; we need a

national effort, not only to gather information, but to devise what they thought about selected aspects of their program.

Required Resources

How are we going to accomplish this broadened agenda with shrinking resources? First, I think if you're in the right area (an area where people want the results of what you're doing) resources are available.

But a second question of concern is critical mass and what resources are required for a viable research and evaluation operation. By my estimates, including salaries, fringe benefits, materials, and everything else, it probably takes about two million dollars. We've got to start working at the state or federal level to explore consortia for these operations. Tom Satterfield's efforts in Mississippi provide one example where districts have banded together and invested in a single unit for research. The idea is to provide sufficient resources to fund a professionally managed and multi-faceted unit, but one which is independent, stands on its own right, and contracts with participating districts to provide services.

Needs for Certification and Training

A final area of priority, one that relates directly to all the others I've mentioned, is training. We need to train people in accounting, management and data processing. These skills are needed if researchers are to communicate with and be useful to decision-makers.

We also need certification procedures to ensure that potential evaluators emerge from school knowing how to really do something.

Again I would recommend having at least five academic areas: accounting, management, data processing, research methodology, and statistics. I'd also demand some full-time research experience for certification, and then for renewal spending a specified amount of teaching time in the classroom (perhaps one semester every five years)--so that administrators keep in touch with what it's like to actually work in a school.

Testing, Measuring, and Evaluating to Support Student Learning and Achievement

Walter Hathaway

Portland, Oregon Public Schools

I believe priorities for research in educational testing and evaluation can be defined from a single overarching question: How can testing, measurement, and evaluation best support student learning and achievement? I offer the following notes as a means of providing a basis for discussion of a number of possible needs:

1. Student Achievement

   a. ways to measure and create an optimal match of instruction and curriculum with individual student needs.

   b. ways to build classroom-up systems of assessment which are useful to the individual teacher, match the curriculum, and can be aggregated for administrative purposes.

   c. ways to improve teachers' assessment, diagnosis, and achievement skills.

   d. ways to help students become self-motivated and self-disciplined.

   e. ways to develop national latent trait model-based item banks for all curricular areas.

   f. improved psychometric theory to identify the relative merits of 1,2, 3 and higher parameter models and the efficacy of particular models for different purposes.

   g. ways to adapt assessment to the needs of special populations.

2. Enrollment

   a. ways to deal with the current enrollment decline.

   b. ways to forecast future trends in enrollment and in student body and community composition.

3. Staff

   a. ways to improve selection and evaluation techniques so as to increase the level of content and teaching skills among the teaching staff, and management skills among administrators.

   b. ways to increase teacher and administrator professionalism, incentive, and performance, and to reduce stress and "burnout" in an increasingly aging staff.

   c. ways to match and modify staff competence to respond to varying cultural and ability groups.

   d. ways to work better with teacher preparation institutions to help them respond better to the real and rapidly changing needs of school districts.

4. Curriculum

   a. ways to help teachers "cover it all" without fragmentation.

   b. ways to respond to new societal needs without overextension or neglect of basics (e.g., multicultural education).

5. Technology

   a. ways to utilize microcomputers to provide tailored testing and instruction.

b. ways to enhance building level test scoring, reporting and utilization through building level computer capability.

6. Community

a. ways to communicate better with the community at both the building and district levels and thus restore confidence in the schools where the loss of this trust is unfounded.

b. ways to involve the community in meaningful and useful ways.

7. Management and Finance

a. ways to gather better information on effectiveness of programs and activities.

b. ways to relate costs to effectiveness

c. ways to support a sense and a reality of participation, autonomy, and control among teaching, administrative, and support staff without losing the ability to act in concert when necessary.

8. Institutional Integrity

a. ways to retrieve or maintain local control over all those functions and resources that must be made common on a statewide or nationwide basis.

Our ability to accomplish these ends will be influenced by three interrelated issues:

--how to help teachers and principals become researchers and evaluators;

--how to foster better collaborative research planning and execution among institutions of higher education, R & E agencies, and state and local school systems; and

--how to develop more effective and timely communication among researchers, especially through electronic networking.

Research as Stimulas Rather than Response

Barbara Chambers

Cleveland, Ohio Public Schools

The problems facing public urban education are numerous and diverse. On the one hand they are typical of those facing public education on the whole; on the other hand they are more severe because each common educational problem is confounded by the additional conditions of poverty, urban flight, and desegregation.

I offer below an array of problems which are especially pressing to urban educators.

1. The disparity in reading achievement between black and white students.
2. The increased number of students who are absent daily from school--especially on the secondary level.
3. The decline in or absence of reliable standards of student performance.
4. The excessive rate of class failure and non-promotion.
5. The decline in the intellectual level of persons entering the teaching profession.
6. The increased number of educational policy decisions either made by non-educators or made by educators without reference to research findings.

Areas of Priority for Research

I believe that the focus of educational research should be on those areas which have direct practical application in the classroom. There is an immediate need to demonstrate to teachers, parents, and

the public that educators (researchers) can identify problem areas in the schools and provide meaningful solutions. The apparent trend of waiting for the media, courts, or politicans to identify public school inadequacies must be reversed.

There are several areas I would suggest as priorities for research and development in evaluation and testing. They are:

1. Student Marks
    a. current practices
    b. validity and reliability
    c. standard setting
    d. recommendations
2. Teacher-Made Tests
    a. validity and reliability
    b. upgrading cognitive level
    c. recommendations
3. Racial Disparity in Reading Achievement
    a. causes
    b. effective practices
    c. recommendations

Above all, I believe it is essential that a variety of vehicles or approaches be developed to deliver the findings of educational research to classroom teachers in a timely, practical manner.

The School, the Home, and Other Educational Resources

Makis Syropoulos

Detroit, Michigan Public Schools

I strongly believe research is urgently needed in a number of areas:

1. Conditions correlated with effective in-service training should be studied. The in-service mission should be analyzed on three levels. First, effective in-service will require strategies for teaching such topics as goals, expectations, and discipline. These strategies have not yet been identified. Second, research is needed that more accurately describes successful school-wide programs aimed at professional growth. Third, in-service must be undertaken concurrently with the programs to help students or teachers whose skills need to be enhanced. It would do little good for an in-service program to successfully achieve school-wide teacher and administrator agreement on particular rules, norms, behaviors, and curricula if, in fact, teachers were incapable of appropriately implementing such agreement, materials or strategies.
2. Research is needed to examine the possible links between in-service education and enhanced teacher productivity. Such links provide the ultimate justification for devoting school system resources to inservice education. Research is needed to observe and describe the dimensions of in-service as experienced by the individual teacher. Most previous

research has focused on particular programs using aggregated individual teachers data to yield group means. Thus, we know very little about the amounts and kinds of in-service education received by individual teachers over specified time periods and about the cumulative effects of in-service education on a teacher's productivity.

3. Research is needed on the personal qualities and social networks of citizens who aspire to be resources and advocates for schools. It is recommended that empirical inquiry focus upon the variety of roles that citizens can perform--advocates for improved education, tutors, or teacher aides.

4. Research is needed on the role of adults as resources for children's learning during television viewing. It is important to consider television watching as a critical social setting--a specific place and time event sequence--which illustrates the interaction between adults and children within the household. Television watching then becomes a catalytic event to clarify and elaborate the child-parent relationship. Parent-child television viewing can be assessed as a potentially positive force in the educational development of the child and a positive resource for the socialization of parent-child relationships. Such research may indicate how technological innovations can become education resources when the context for parent-child learning interactions is understood.

5. Research is needed to investigate the similarities/differences among effective teaching processes that occur in the home and in the classroom. There is some evidence that specific teaching practices are of primary importance to children's learning, whether these practices occur in the classroom between teacher and child or whether they occur in the home between parents and child. There is also some evidence that these processes can be mutually reinforcing. One important result of such research may be an illustration of how parents and teachers can collaborate in stimulating and reinforcing children's learning. School policies to support this collaboration can be possible consequences of such research.
6. Research is needed to investigate the role(s) of teacher centers on complementary and/or initiating agencies in promoting more effective schools. The inherent, collaborative policy board of federally-sponsored teacher centers may serve as a model for instructional enhancement at the school level.
7. Research is needed in the area of management practices at the local school level that are necessary to help education professionals improve basic skills instruction.
8. Research is needed to study how effective schools got to be that way. This research will likely improve managerial practices at the local school level. Developmental research may expose the critical managerial decisions, actions, and role perceptions that culminate in an effective school.

9. Research is needed in the area of computer literacy for both elementary and secondary levels. If research is not possible at this time, maybe a commission should be appointed to come forth with some curriculum guidelines to help local school districts select computer literacy curricula for their students.
10. Research is need in the area of computer software. Like computer literacy, some guidelines on how to select software for the school would be very helpful.

Evaluation and Accounting, Cost Analysis, and Management

James De Gracie

Mesa, Arizona Public Schools

Evaluation at the local level has evolved over the last twenty years or so toward more multifaceted approaches. No longer are we just the testing department. In the majority of studies we currently conduct, we are called upon to do at least cost analyses, if not cost effectiveness analyses. In our district, we're also involved in the management studies areas and are faced with problems such as "Are our food services cost effective?" or "Can we transport the students over the 200 square miles of the district in a more efficient manner while maintaining maximum safety?" We're actually doing more of these kinds of studies than dealing with testing, although we still use standardized tests to look at program effects. We therefore need to develop better techniques and assure that our personnel are well-trained in accounting, cost analyses, and management as well as in educational measurement.

A second area of high priority is the high school. I'm not sure of the impetus, whether it is the effective schools studies or the major studies that are going on in the high schools, but our superintendents and deputy superintendents are all emphasizing the high school. They seem to be saying, "Okay, we've just monitored and assessed the elementary schools; it's time to do the same to or for the high schools." This sentiment seems common to all our districts. As a result of this interest, a number of districts are being called upon to develop tests in a whole range of curricular areas: Algebra I,

French I, Algebra II, and so on. This requires a massive effort at the local level, an effort that really surpasses available resources. Sharing among districts helps some, but we really need more help.

Another area of high concern is microcomputers. There seems to be a need, or rather a desire, to have an Apple on every desk, but we really need to look more closely at the costs and benefits of micros. For example, in our district we're trying some programs with computer-managed instruction, some with computer-assisted instruction, and some tutorial. But despite the glowing reports from salesmen, I'm not sure any are really producing more knowledgeable students or less paperwork for teachers. We may need to develop our own micro software to match the district's goals and objectives and the specific needs of teachers.

Costs will be a big issue in microcomputers. Our deputy superintendent wants microcomputers in the classroom as long as we can guarantee him they're going to save time or personnel and improve achievement, and that they're cost-effective. We need to demonstrate that a micro is better than a textbook, otherwise our district won't buy them for management or CAI.

Data-Based and Humanistic Investigation: Issues that Would Profit From a Merger

Eric F. Gardner

Syracuse University

A response to the request to comment on critical R&D needs in educational testing and evaluation can profitably be preceeded by three general statements:

1. There are few, if any, measurement or evaluation problems in the 1980's that are not rooted in the 1970's.
2. The major problems facing the measurement field in attempting to fulfill the future needs of education will be not technical but political (this comment is not made to minimize the large number of existing technical problems needing research). This point was illustrated recently by Lorrie Shepard in her 1983 NCME presidential address, "The Role of Measurement in Educational Policy: Lessons from the Identification of Learning Disabilities."

Among the various political manifestations (which in turn create additional political problems) has been the tendency of educators to become polarized and to form two often opposing camps. One group has a major interest in the solution of problems and the making of decisions through the systematic use of data. The second has emphasized the humanistic aspects of education and tends to perceive the first group as big, impersonal organizations whose members care little about individual human beings. Worse still, some attempts of individuals in the second camp to use measuring tools, or to

participate in the construction of one, changes the way she/he is perceived by the members of her/his own group.

This bickering among the professionals tends not only to confuse the ultimate decision-makers--parents, lay citizens, school boards, and legislators--but to undermine the credibility of professional education and professional educators. It is easy to attach blame for such a situation to any group one's personal biases. However, we professionals in the measurement field should be credited with the major blame. As David Orr so aptly put it at a Phi Delta Kappa symposium:

> The [test] user has been conditioned to accept the superiority of numbers as part of the magic of modern technology. Inevitably the fallibility of test scores surfaces, and the disenchanted user becomes another critic. But this result stems not so much from the technical deficiencies of measurement as from the failure of communication about what to expect and how to use the measures--a 'political' problem.

3. Evaluation studies and research are likely to be ineffective unless the specific interested publics and their objectives and interests are identified and addressed directly. There are a variety of publics (general public, legislators, school boards, administrators, teachers, parents, and pupils) concerned with the outcomes of education. These groups need an understanding of specific measurement concepts and the skills to interpret teacher and pupil output. The professional in the evaluation and measurement field must identify these publics and, with their cooperation, develop the necessary communication lines and techniques needed to

provide the data essential in answering their questions about education. Such an approach with shared responsibility is essential if the field of measurement is to fulfill its responsibilities in the 1980's. Some steps have been undertaken by NIE, Phi Delta Kappa, AERA, APA, NCME, ETS, and other organizations. However, relative to the magnitude of the task, we have hardly begun.

A Few Problems Needing Research

Within the context of the previous admonitions, the following issues are a few of those needing considerable research.

1. Validity. The word validity has tended to have different meanings depending upon the situation and the person using the word. The most recent proposal and definition of one aspect of validity was Robert Ebel's suggestion of "intrinsic rational validity" for ability tests published in the Summer 1983 issue of Educational Measurement: Issues and Practice. Research is needed which might begin with a definition of construct validity under which all other types would be subsumed and which would present clear, unambiguous definitions of other types of test validity emphasizing the importance of the purpose for which the test is to be used.
2. Study of achievement in terms of cognitive psychology. Over 25 years ago, Ann Anastasi lamented that psychologists and psychometricians were diverging as they addressed common problems and were reaching the point where they could hardly communicate. This divergence is continuing and needs attention.

3. Research leading to a theory which defines an underlying continuum for prerequisite variables. The identification of variables which relate ability (especially in generalized situations) to later performance are needed.
4. Research leading to the measurement of a number of affective variables such as stress and attitude toward teaching as a career. The most successful measurement instruments have been in the cognitive domain. In spite of the difficulty in constructing measures of affective variables, much work is needed in this area.
5. Research on the measurement of values using a broader approach than Kohlberg. Values, years ago, were not considered a relevant part of psychology. They are now considered important by many psychologists and need to be defined and measured.
6. Research on methods of vertical equating. Studies relating the current methods (equipercentile, regression, and latent trait) as well as exploring new approaches are needed to determine the best approximations and the associated error for specific situations.
7. Research on the computer. Research leading to more effective use of the computer for test administration, immediate feedback, test construction, and information dissemination is in its embryonic stages.
8. Interpretation of modified standardized tests. Empirical research is needed which shows what the effect would be on

the norms by the addition or substitution of a few local items to a standardized achievement test.

9. Norm-referenced versus criterion-referenced tests. Research is needed in order to examine whether norm-referenced and criterion-referenced tests are overlapping subsets of a broader continuum.

10. Research on the type of construction and sampling of items needed to most completely represent a domain. Some tests in which each item focuses on a very narrow objective omit coverage of more important broader objectives which are crucial to the adequate measurement of the domain.

11. Research on effective ways of measuring change. In spite of the extensive research on statistical issues involved in measuring change, there are a number of unanswered questions related to the kinds of instruments and techniques needed to demonstrate effective improvement of individual pupils and groups.

Issues in Bilingual Education

Jose Vasquez

Hunter College of CUNY

Over the past 15 years, bilingual education programs have been designed, promoted, and litigated primarily as a measure to correct "English language deficiencies." Bilingual instruction has received support from federal and state agencies because of the widespread assumption that the inability to understand and speak English fluently accounts for the low academic achievement of many ethnolinguistic minority students.

In the United States today, bilingual schooling is perceived as a means to remedy inequality in educational opportunity for students who speak no English or are of limited English proficiency (NEP/LEP). The argument is that, if a child comes to school with a home language other than English--the single language of instruction in the nation's schools--opportunity to learn is denied primarily because the child cannot understand what is going on, cannot communicate with the teacher and others in the classroom, and therefore is rendered incapable of participating competently in instructional activity,

More than most educational innovations, bilingual public education operates in a context that is charged with controversy. There are conflicting views about the meaning of equality of educational opportunity, the impact of minority language group isolation, and whether or not societal institutions (most particularly government and schools) are obligated to foster and preserve cultural and ethnolinguistic heterogeneity. In addition, the involvement of

politically oriented groups has sometimes led to a confusion of political motivation and educational justification in the formulation of bilingual education policy.

Bilingual schooling is an idea striving to find effective forms of expression, and an educational movement attempting to shape its own course. One of the major reasons that bilingual schooling has not yet been grounded in empirically validated findings is the lack of a firm research base that provides solid evidence for educational decision-making. Research is complex and expensive, especially the kind required to cover the diverse and extremely complicated variables involved in bilingual education.

Without the findings from a sufficient body of research endeavors, both the advocates and the opponents of bilingual education have been duelling in a fog. Each side contends that it is right; neither has the evidence to prove it.

Another area that has been problematic for resarch is what exactly to study: how can the effectiveness of bilingual education be meaured, and in relation to what? What is the relation between measures of program effectiveness and measures of student achievement?

Unfortunately, most of the research on bilingual education in the United States has mostly looked at the overall consequences for the student, but findings have been in global terms. The reason is that the research has been in the form of quantitative program evaluations of reading and math achievement, in the absence of qualitative measures as well.

As a consequence of this state of affairs, in 1978 Congress requested specific information to determine the needs of bilingual education in the United States. In Title VII of the Education Amendments of 1978, Congress directed the Secretary of Education to "develop a research program for bilingual education." This agenda is comprised of a series of studies organized into three general research categories.

A. Assessment of national needs for bilingual education

B. Improvement in the effectiveness of services to students

C. Improvement of Title VII program management and operations

The 1978 amendment of Title VII that mandated research was a step in the right direction. For 15 years, bilingual educators have been scaling a mountain--the mountain of ethnocentrism, of bureaucracy, of apathy, of entrenched educational ideas. Our progress has certainly not been uphill all the way. At times we have taken wrong paths, come up against steep cliffs, almost lost our footing. This was to be expected, since we did not have the map and compass that research could have provided to get our bearings. Yet we've persevered, and the longer we've climbed, the better we've become at judging the trail we should follow.

Unquestionably, there have been two major problems in designing effective bilingual instructional programs for children who speak no English or are of limited English proficiency. First, the children in need of such services must be identified; then, after they have received the services, a decision must be reached about their placement in the educational system. The everlasting entry/exit

dilemma is basically one of determining whether the student is ready to participate successfully in an educational program with all instruction being given in English. Transition from the minority language to that of the majority is still viewed as being a panacea. This perception, overlooking the value of maintenance and the appeal of enrichment, relegates bilingual education to a remedial role. The mountain of its potential is diminished to a molehill. Because of this narrow perspective, the results of English achievement tests are overly relied upon as an indicator for entry and exit.

The Significant Bilingual Instructional Features Study (SBIF), which is a product of the Part C Research Agenda mentioned above, has revealed instructional practices for the non-English/limited English-speaking student population. This study, built on findings from research on effective instruction, has established a relationship between effective practices and student participation and achievement, through a time-on-task variable referred to as Academic Learning Time (ALT). The classroom behavior(s) of both teachers and students have been observed and analyzed in the SBIF Study. This has provided insight into the student diagnosis, and the entry/exit dilemma, because it documents the extent to which students can attend to what is going on in the classroom. The assumption is that if non-English or limited English-speaking students can attend to instruction, then they are able to participate meaningfully in class. These are precisely the procedures needed to determine the extent to which students are participating in a a given instructional setting. Therefore, we must explore the feasibility of employing these procedures for student identification.

Because I have been immediately involved in the SBIF Study, I have centered my remarks on it. However, I recognize and applaud the entire body of Part C Research. These studies have contributed to the identification of solutions for a wide variety of academic problems; they have revealed ways in which bilingual instruction can be made more effective; and they have given shape to what had heretofore been an amorphous educational movement. Moreover, they have a resonance that extends beyond the boundaries of communities in which they were done. It is my recommendation that the findings from these studies be synthesized, summarized, and brought to the attention of practitioners and policy makers throughout the nation. The importance of this research to bilingual education is inestimable; but like bilingualism itself, it can be of great value in the advancement or improvement of any educational discipline. Yet I am positive that a major decline in funding for the research efforts of the Title VII network could stop us in our tracks. The policy makers must hear us when we say, "We have come so far; let us press on."