R&D PRIORITIES FOR EDUCATIONAL TESTING AND EVALUATION: THE TESTIMONY OF THE CRESST NATIONAL FACULTY

CSE Technical Report 304

Joan L. Herman, Editor

UCLA Center for Research on Evaluation,
Standards, and Student Testing

September, 1989

The research reported herein was conducted with partial support from the U.S. Department of Education, Office of Educational Research and Improvement, pursuant to Grant No. G0086-003. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

Table of Contents

Introduction

The National Faculty of the Center for Research on Evaluation, Standards and Student Testing (CRESST) is an affiliated network of practitioners, policymakers and researchers who have interests in educational testing and evaluation. The group was constituted to support CRESST in meeting its mission by assisting in needs sensing, project planning, review and dissemination. The group has been convened annually in conjunction with the annual meeting of the American Educational Research Association.

The 1989 meeting of the faculty, held in San Francisco, was convened expressly to serve a needs sensing role. Faculty were invited to present testimony on what they viewed as the most pressing research and policy issues in the fields of testing, evaluation and standards. The document which follows expresses these views, highlighting important problems for a national R&D agenda.

While the collected views represent a diverse group, including state and local level policymakers and practitioners, test publishers, professional organizations, and other researchers and R&D providers, they nonetheless show considerable agreement on priority areas. Chief among these is the need to develop alternatives to currently pervasive standardized, multiple choice tests. Repeatedly expressed was the need for meaningful, performance-based assessment of higher level skills which, in the words of Beverly Anderson, Educational Commission of the States, avoids "a piecemeal approach to education and assessment" and instead is "based on an authentic performance, uses relevant content, and employs valid criteria in a sound scoring process." Todd Endo, Fairfax County Schools, suggests that building up from what teachers are currently doing is a good starting place for developing such alternative measures. David Berliner, Arizona State University, Gerald Bracey, Cherry Creek Schools, and Lynn Winters, Palos Verdes Schools, raise some of the technical issues which surround the development of sound alternative assessments, notably new approaches to validity, reliability, and comparability.

Beyond the need for alternatives which better capture significant, higher level skills, several also noted the need for process oriented measures which could help teachers diagnose and better facilitate students' learning. Merlin Wittrock, UCLA, for example, speaks to the research literature which might ground such new measures and George Malo, Tennessee State Department of Education highlights some of the multiple assessments which could help teachers. Berliner concurs with Wittrock in the need for process measures and advocates alternatives in the affective domain as well.

Paralleling the call for alternative types of assessment was a relatively frequent rejoinder for better quality indicators. Sharon Robinson, National Education Association, asks for ways to help schools better document what they are doing and accomplishing, while Lori Orum, National Council of La Raza, echoes an interest in non-standard indicators and qualitative assessments of school success. Sharon Johnson Lewis, Detroit Public Schools, adds a particular emphasis on indicators that will help us to identify underlying causes in educational process and context and to understand why things are as they are and how to make them better. Stephanie Burton, National Computer Systems, proposes the need for a common taxonomy, an "educational esperanto" for describing educational experiences and outcomes.

What standards should guide the development of such quality indicators? Methods for reaching concensus on such standards was identified as a continuing problem (Robinson) as was the identification of the skills and social predispositions valued by various constituencies (Endo and Orum).

Also signaling reservations about existing practice, a third common theme focuses on the impact of current accountability practices. Common was a concern for the effects of testing mandates on curriculum, teaching, and instruction, expressed by Endo, Robinson, Winters. Highlighting potential positive impacts, John Keene, National Computer Systems, recommends research and where and how assessment mandates improve educational quality. Gary Estes, Far West Regional Laboratory, and George Malo advise research and development which will moderate some current shortcomings, namely how to build in sufficient local autonomy and how to increase the efficiency of testing.

Frequent also was the call for research and development efforts which could help improve the utility and intelligent use of testing and evaluation information. Keene stresses the importance of "demystifying test scores" for the many intended users of assessment results -- educators, legislators, the public. Bracey, Malo, Robinson, and Orum likewise see a significant need for training, while Gary Williamson, Greensboro Public Schools, Tom Kerins, Illinois State Department of Education, and Winters address themselves to the need for better ways to organize, store, and report information for school improvement.

A concern for more effective communication of evaluation results and the promising practices they imply also was expressed. Keene, Robinson, and Williamson particularly ask that the field more actively synthesize what it has collectively learned and better disseminate its practical implications. James Olsen, WICAT, Estes, and Johnson-Lewis remind us to both continue to scrutinize the impact of various reforms and innovations and broadly communicate such findings.

Finally, testimony givers raised various technical questions that merit attention. Among these are issues in developing content hierarchies on which to scale performance; test equating; and statistical methods for assessing growth.

The testimony follows. We at CRESST appreciate the thought and effort that this testimony represents.

Beverly Anderson
Educational Commission of the States

I come from the perspective of state leaders with whom we work at the Educational Commission of the States (ECS). As some of you know, we work mainly with governors and legislators and other state leaders through ECS. One of the things that we're finding is that these folks are very concerned about the status of students' ability to solve problems and to think well about important issues. There is growing concern that students are leaving schools without really being able to wrestle with difficult situations.

As we've looked into the problem more and more, we see that the whole education system tends toward a kind of a piecemeal assortment accumulation of information and discrete skills. What we're looking for are ways to shift the balance of the education provided by schools so that it's not quite so lopsided. There is instruction, there's goal setting, and there's assessment, and they all must support a meaningful total, total that fully integrates understanding. For example, students tend to get bits and pieces of information about English in one class and then they go to science and math. But they never see how all of those pieces from those courses fit together to deal with a real life problem, such as an environmental issue in their community or a problem in an on-the-job situation. If anything, it seems that we're moving more and more toward that kind of piecemeal education.

As we've looked at it from an assessment perspective, we find similar lopsided attention to a piecemeal approach. Look at the research and types of assessment which are going on. We, as a research community, can't afford to have that kind of bias coming through in the research we're doing.

While I'm sure there are many ways to think about this problem of piecemeal systems, one approach we've been trying to encourage is authentic measurement. Authentic measurement is measurement that assesses meaningful performance, uses relevant content, and employs valid criteria, in a sound scoring process. Let me just give you some examples of what I mean: We need measurement where the task is an authentic, meaningful performance. That is, the task is something that you want students to do well in the future. So the very measurement task itself is something you want students to do well. For example, you want them to write well, or you want them to engage in good discussion, or you want them to be able to enjoy a good book or discuss abstract concepts. Contrast this with typical tests now. Do you want students for the rest of their lives going around answering four-choice multiple choice questions? That's not an authentic performance. We need more measurement that is based on an authentic performance.

Secondly, in terms of measurement that's based on authentic content, a major problem we see here is that the most current measurement tasks are not inter-disciplinary. Yet most of the situations people face in life are inter-disciplinary. If your task is to build a fence, it's not just a mathematical problem. If you're going to build a good fence, you have to know something about weather conditions, you have to know something about the preferences of the person for whom you're building the fence, and then of course you need the mathematical skills. Current measurement is not sufficiently based on authentic content.

A third aspect is that we need situations where the measurement strategies include known, authentic criteria. The problem that we see here is that the students tend to get their results back in terms of whether something is right. Such as, we got 90% of the right answers. What is not made clear is what exactly are the standards of judgment by which it was decided what was right? In writing assessment where papers are analytically scored, there is usually much more reporting of what the criteria are. I think we need to keep emphasizing such reporting in our measurement to highlight what the criteria are, as well as how well someone measured up to that criteria. Unfortunately, now we only emphasize how well students measured up, and users simply don't even understand what the criteria is. So the measurements become meaningless.

Finally, I'll just comment on authentic scoring. Here what we lack are approaches where the scoring itself involves people and not just machines, people who understand the criteria. In real life, people's performances are always being judged by other people. What we're looking for is a measurement process where, and again writing assessment is a good example, people come together and reach a consensus on what the criteria are, and then they apply that judgment based on those standards. In summary, I think these kinds of authentic measurement issues need to be emphasized in the coming research agenda. Writing assessment currently is the best example of these concerns, but we need to expand the approach into additional areas, e.g., judging students' ability to engage in an effective discussion of important content, to look at social dilemmas, to work cooperatively, and other important tasks that represent this more authentic instruction and learning.

Gary D. Estes
Far West Regional Educational Laboratory

Assessment and Evaluation Problems and Researchable Issues

Problem #1. Evaluating School Improvements Efforts

A variety of approaches, systems, and agencies are engaged in school and program improvement. Many of these have a common feature of promoting and supporting the use of research on effective instruction, curriculum, and organizational practices. But there is not strong evidence (a) that these approaches are working well in those schools or programs most needing to improve and (b) that those who undertake improvement actually improve beyond what would have been expected without undertaking the improvement efforts.

Researchable Issues.

What is the distribution of school and program improvement efforts across categories of schools and programs where the categories include level of achievement before improvement is undertaken, concentration of poor students, inadequate leadership or staff? What are the changes in achievement, what is the evidence that significant changes are made, and are there indicators of the breadth of impact that goes beyond achievement test scores, dropout rates, and attendance to include factors such as "success in subsequent grades, introduction to and success in more advanced curriculum for more students, etc.?"

Problem #2. School Flexibility and Accountability

Several states, e.g., Hawaii, Washington, and Oregon, are attempting to give schools more flexibility in terms of rules and regulations while focusing accountability on student outcomes. It is not clear what the tie or relation is or should be between these two objectives, i.e., to grant flexibility and increase accountability. The effect of legislation and activity in this area does not seem well documented in terms of either the degree to which schools exercise the flexibility by changing significantly or undergoing "reform" efforts. Also, there is little information to support what schools should be ultimately held accountable for in terms of student outcomes.

Researchable Issues

What is the effect of the legislation and programs that attempt reform through increased flexibility for school level management and increased accountability? What support systems are needed to insure schools make meaningful changes? What do schools do that they could or would not have done otherwise? What assumptions underlie the polices? For example, what do they assume about local capacity to change?

Problem #3. Early Childhood Programs

Increased attention is being given to early childhood education and care programs. Emphasis in this area derives from factors such as the increasing number and proportion of children living in property and the increases in households in which both parents work or which are headed by a single parent. Head Start programs represent perhaps the best information source on approaches and instruments that can be used to evaluate early childhood programs. But new programs such as the federal Even Start program have approaches that are indirectly attempting to benefit children--such as through parent literacy training. It is unclear what methods, instruments, and approaches will be useful in evaluating and assessing the effectiveness, costs, and benefits of these programs.

Researchable Issues.

There is a need for evaluation guidelines, approaches, examples, even models, that can efficiently evaluate "intergenerational" programs such as the new Even Start program. What approaches can local agencies use to examine their efforts? What instruments or indicators can be used?

Problem #4. Measures that Exhibit Instructional and Curricular Validity

To date, the majority of measures of student achievement from the instruction and curriculum they receive in school comes from standardized, multiple-choice tests. These instruments have been useful over the years in providing indications of student progress in areas commonly emphasized across the various textbooks used. But as increased attention has been give to student outcomes as reflected in these measures, they have moved from the place of sampling student performance in a larger domain that was to be represented by the tests to having student's instruction being overly driven to emphasize performance on these measures. The problem is not only in the format of the tests, but includes other issues such as test item security or degree of exposure. It appears that accountability is not on the decline and that whether right or wrong, schools and staff will need to demonstrate evidence of student achievement on some measure. Writing assessments are probably one of the better examples of an assessment where "teaching to the assessment" minimizes instructional distortion.

Researchable Issues.

What assessment approaches can be used by schools and teachers that are both useful for accountability purposes and which exhibit instructional and curricular validity? That is, the tasks required by the tests look like those that students are required to do in normal instruction and learning. How can technology and alternative assessment approaches be used to increase the match between instructional and assessment tasks? How should differences across subjects, e.g., reading, math, social studies, music, art, be handled in assessment and accountability?

Problem #5. Classroom Teacher Assessments

Although standardized tests receive most of the attention, from a student's viewpoint, the majority of testing or assessment is that which their teacher constructs or uses. Research and training by my colleagues at NWREL, e.g., Rick Stiggins, clearly documents this problem and offers ways to improve the classroom teacher's assessment competencies. But what remains a problem is how to have a greater impact in changing classroom teacher practices.

Researchable Issues

What are options for assisting teachers to improve their assessment? Can technology be used effectively such that teachers could improve their skills and practices? What are some options and pitfalls in displacing or supplementing standardized assessments with classroom assessments?

Lori S. Orum, EXCEL Project Director
Sylvia A. Alatorre Alva, Ph.D., Research Specialist

National Council of La Raza

In support of the Center for Research on Evaluation, Standards, and Student Testing's work in the area of testing and evaluation, I am pleased that you have invited us to identify the most pressing research and policy issues that affect the evaluation of our project -- Project EXCEL. We feel very fortunate to have the support and technical assistance of CRESST and look forward to continuing a collaborative relationship around some of the issues I have highlighted.

As you know, the National Council of La Raza's Project EXCEL (Excellence in Community Educational Leadership) is a national effort to improve the educational status of Hispanic children, youth, and adults by working with community-based agencies to demonstrate the effectiveness of several educational models. The Council firmly believes that community-based agencies are valuable partners in efforts to improve the quality of education available to Hispanics. However, evaluating educational programs that operate from community settings is a largely uncharted area of research and policy, with its own unique set of challenges.

For example, by and large, the staff members in CBOs have not had formal training in the area of testing, assessment, and program evaluation, nor do many of the agencies that we work with have the resources to hire a consultant or evaluator. Consequently, they require extensive training on how to collect and interpret evaluation data, as well as assistance in monitoring the implementation of their programs. While the Council plays an ongoing role in monitoring and evaluating the demonstration sites, ultimately, the quality and validity of the information that we obtain for evaluation purposes depend upon what took place at the local sites. Thus, it is important to develop training materials on how to implement and evaluate programs that are written for a community-based audience that is not necessarily trained in assessment and program evaluation and to provide a framework for why evaluation is important and what can be learned from evaluation results.

An additional area deserving more in-depth study involves the over-reliance on standardized tests to measure academic achievement and progress. This has long been a topic of debate and controversy, because tests are often not designed or normed for the population served. While there are serious shortcomings in how test results are used to inform research and educational policies that affect Hispanics, we cannot ignore the need to assess the performance of individual students or the effectiveness of entire programs.

In response to these concerns, we recently embarked on an exciting writing assessment project, in conjunction with the Educational Testing Service, building upon the work we have been doing with CSE/CRESST, to develop a procedure to assess the writing ability of students and parents in our after-school enrichment and adult literacy programs. Using Holistic scoring, we are in the process of developing a scoring procedure for limited-English proficient parents in our adult-literacy programs. This, of course, raises an entirely different set of issues. As you are well aware, the methodology for scoring writing samples cannot just simply be transferred from English for use with adults who are limited proficient in English. With non-native English populations, special consideration must be given to developing a scoring criterion that measures not only writing ability but also the acquisition of a second language.

The National Council of La Raza believes that the following areas are important -- and inadequately addressed areas for additional focused study:

- In addition to standardized test scores and school grades, what non-standardized measures can be used as indicators of school success?

- How can qualitative assessment best be used in community-based programs?

- Beyond test scores and grades, what skills can be developed in children that will allow them to cope with the educational and social demands that put them at high-risk of dropping out of school?

- How can holistically-scored writing assessment be used to measure literacy developed (in English and Spanish) in youth and adult learners?

- With our adult literacy program, we have found that there is very little baseline information on the literacy skills of adults who are not native English speakers. As an example, the Test of Adult Basic Education (TABE), one of the most frequently used measures of adult literacy, is not available in Spanish and has been normed for this population.

In closing, we have really enjoyed working with the staff at the CSE/CRESST and have benefitted greatly from the Center's expertise. We look forward to continuing our collaborative relationship.

Sharon Robinson
National Education Association

I have outlined my notions about critical issues for the research agenda or mission for the Center, and it was quite fun to do. I think of the work that is currently going on at the Center and in the field. It's inspiring. It's very interesting, but we don't know enough about it.

I would begin my agenda by asking that we return to greater concern for theory development, attend to the synthesis of work that is ongoing, and develop a strong knowledge base about assessment.

Then, once accumulated, the next step is getting that knowledge out, particularly to practitioners, and from there I would move into the second notion of what I would call "industry accountability." There are standards for all the technical requirements of putting together a test, but while those requirements get published and discussed at meetings such as this, I find that their operational reality is very moot. It would be really interesting to have some method by which the standards throughout the measurement industry are assessed and shared, for all of us to look at and be aware of.

Other folks have discussed the impact of the various assessment laws throughout the states and at the local level on actual classroom practice. I think that the story needs to be told in a very organized and disciplined way. We need research into the legal implications of testing, in terms of truth in testing type legislation, but also in terms of the impact on curriculum. When we think about the impact of testing on the curriculum, I have to raise again my concern about pushing tests further and further down into the age levels. The testing of preschool or early childhood education children is something I see on the one hand while on the other you're telling me we're getting better at testing. Illinois and California are going to show us the way, but these states are still testing kindergarten students, who don't know how to hold a #2 pencil yet.

I think that informing public policy makers about different approaches to accountability is a very important priority for the center. Until the policy makers and the politicians become convinced that the educational institution can in fact report on its activities and its progress in ways other than highly standardized, norm-referenced measures, then we're doomed to that technology.

In the area of evaluation, I would urge more work in helping schools document what they're doing, particularly documentation at a school site level. Right now we have a lot of data district-wide, but we are not capturing some very important, rich improvements that are going on in various school-based efforts that can not be captured in standardized tests. And that is why they are rich. We really do need to know how to do that quickly because teachers are producing a lot of important results that we don't know how to talk about very well. Furthermore, we don't even know how to say what needs to happen for this to happen more, and spread the potential, or make the capacity even greater.

Finally let me talk about standards. I don't know what a research community can do about this, but we don't have any consensus at all about what it is we're trying to achieve in schools. Or what the standard is, so that we'll know if we're getting there. I fear that as long as this problem persists, we will have the riding tide of mediocrity about every 20 years or worse. I think we've got to come to grips with some methodology for reaching consensus, and school districts have got to understand the need for expressing a consensus view about standards, what it is they're trying to do, and the performance level that will be achieved when that happens. That's imponderable, I worry about it all the time, maybe CRESST can worry with me. Thank you.

Gerald W. Bracey, Ph.D.
Director of Research and Evaluation
Cherry Creek Schools

For me, most of the salient issues revolve around a single question: how can we quickly but carefully develop alternatives to standardized tests whose inadequacies to perform in the arenas of both instruction and accountability are so painfully evident?

The issues of reliability and validity loom large when one starts thinking about alternatives. Our conceptions of reliability, by and large, assume the peculiar technology of norm-referenced tests, despite much work in criterion-referenced testing. We may have to reformulate both constructs to make them adequate to judging performances. For example, if a professional athlete has an off-day, that is what we say. If he or she is highly variable over time, we call him or her inconsistent. In both instances, the locus of the variable performance is in the performer. In psychology and education, we say that the test is unreliable.

A second issue of concern is that of standards. What constitutes an adequate performance on a test and by what criteria?

At a practical level, though, the largest issue for me concerns staff development. Although teachers spend some 20% of their time in assessment related activities, they are not trained in how to conduct such assessments well. Indeed, in preservice training, classroom assessment is likely either ignored or discouraged in favor of standardized tests constructed by "experts." Most alternatives to standardized tests require extensive and critical judgments about performance by teachers. The staff development implications of this are enormous, but research in effective staff development is meager.

Finally, as I feel I have shown, the so-called basic skills tested by standardized tests are precisely that-so-called. A research program to determine if there are universal basics, or if situated cognition means that cognitive development is idiosyncratic, should be a fruitful program. Only Jean Piaget has dealt with this in a systematic way from birth onward.

Todd Endo
Fairfax County Public Schools

I am not here as a dispassionate, disinterested researcher. I am not here even as a disinterested practitioner. I have biases in the testing and evaluation arena, and they are that we in fact test too much, and that we use tests too much. A number of people have already made these points.

In thinking about what should be on the research agenda, I posed two questions for myself: What is it that I want to know? Also, since I am not responsible for CRESST, if I were the director of CRESST, what would I do? It is probably impolitic, but one of the things that I would suggest is that CRESST establish a vision of where we want to be in this arena of standards assessment testing 20 years from now and help us get there. If that is thought of as not the proper approach for a research center, so be it. I think we're in a political arena, as much as a knowledge arena, and we ought to be looking at where we want to be in 2010 in the arena of student assessment.

Since most everything else that I'd want to say really already has been said in one way or another, I want to reinforce what to me are the important issues for the research agenda: I'd start with standards. There's been a lot of work on standards, perhaps even sufficient research, but CRESST could synthesize it, publicize it, and put testing in a much broader framework.

A second important research area is what the various constituencies for education think are important student outcomes. There has been a lot done in this area, but perhaps there needs to be more. What do employers really think students or people coming into the employment arena need to be able to do; what kinds of skills do they need? What do teachers say? What do students say? What do colleges say? Start with those desirable outcomes, and then work from there towards what do you do about it, how you assess it.

A lot of the rhetoric today starts with existing tests, and then goes to how you use them. I'd favor more balance. If we are going to talk about the appropriate use of norm-referenced tests, and how we can help teachers use norm-referenced tests better, I think that's fine, as long as we spend an equal amount of time, and maybe more, on the appropriate use of essays. I don't hear too many people talking about appropriate use of essays. Take any other assessment means besides norm-referenced tests, whether it's a science project or a research paper, or skills one would need to be a good employee--and work through the same kinds of issues we're dealing with norm-referenced tests, and I think we'd have a little better balance.

Perhaps we can learn something from the National Board for Professional Teaching Standards, a topic which gets heavy coverage in Education Week this week. Maybe we need to have something like a national board for student achievement standards. What the teaching standards board has done is publish a whole set of standards that they think a good, experienced, excellent teacher ought to have, and then they think about how to go about assessing whether a teacher has those. The first stage has been laying out these standards.

A second topic in Education Week is school districts and their testing programs. The one that is in this issue is Prince George's County, Maryland, which is a nearby school system to me, and many of my colleagues are there. One of the things that research could do, and again this has been said, is really look at the impact of high stakes testing on teaching and learning in schools. Take Prince George's County, a good study of programs like this would confirm whether Murphy is right: "If you walk through schools and look at faces," he says, "you'll see excitement for learning that was not there five years ago." That's more important to him than CAT scores. On the other hand, a teacher says, "Murphy is proud of climbing test scores, but I don't see that reflected in improvements in the way the students use the English language." It goes on and on, quote after quote, of the advocates and the dissenters, on what have been the consequences in terms of teaching and learning in a system that has put relatively high stakes on testing.

The last thing I want to say is, reinforcing what Sharon Robinson said, I'd start not just with standards, but I'd start with that practitioner out there and examine practices closely. Start with the teacher. Sharon said, look at schools--I would support that. How do schools evaluate student learning? And what are the good exemplars out there, how do they go about it? In a school district I find that the teachers do not use tests much, unless they are forced to do so. There are good teachers out there who are doing a good job of assessing student learning. Go into classrooms: how does a good teacher assess student learning? Maybe we can build from there in terms of better assessment procedures that other teachers might benefit from. Attention to such assessment should be in addition to, not in place of, further inquiries about external tests.

We need both types of assessment, but thus far we have not mentioned a proper balance in the way we examine testing, standards, and evaluation issues. We're focusing too much on tests, pencil-paper tests, external tests that a teacher as a technician in the classroom or the principal might use. We need to consider alternatives. Let me end there.

Sharon Johnson-Lewis
Detroit Public Schools

Research for the 21st century should not focus so much on problems, but on a clear delineation of what makes a difference and why, and how this information can be successfully disseminated by objective spokespersons to the necessary policy makers.

Research should not focus on how many students are dropping out of school--but what are the ingredients of successful programs for preventing urban White, Black, Hispanic and other youth from leaving school early? How do you successfully turn significant numbers of 13- to 16-year-old students around? What does the educational environment look like? What is the extent of parental involvement? How do these programs differ from what is traditionally tried in schools now?

Research should not focus on the fact that urban school districts are not successfully preparing students for work or higher education--but how are school districts successfully working with businesses and institutes of higher education? What do these long lasting partnerships look like? What was the role of business in defining their responsibilities and the responsibilities of the school? What was the role of the education system in defining these responsibilities? How are those important links formed so that both groups feel that they are winners?

Research should not focus on the fact that parental involvement is important and on a decline-- but what are the characteristics of a successful parenting program? What were the activities that led to parental participation? What did the parents do differently? And finally, what were the variables that made these programs successful in comparison to traditional programs?

Research should not simply state that preschool programs are an effective way of curbing the dropout problem--but what are the characteristics of a successful preschool program that focuses on improved student learning, and what are the consequences of not having the program properly implemented? How do those characteristics of a successful preschool program differ from a preschool program that does not have the same success?

Research should not simply state that test taking skills are important--but distinguish what those test taking skills are and identify successful programs that operate within the public schools. Who teaches them? What type of teacher training is necessary? And how many hours of student learning time is involved?

Research should not simply focus on the fact that minorities and women generally score lower on college entrance exams--but what are the successful strategies that help improve those test scores? What are the learning conditions that consistently make a difference? What teaching methods are used? How much student learning time is required? And what should be emphasized at each grade level with successful teaching strategies included?

Research should not just point out that poor student attendance has a significant impact on student achievement--but what strategies have consistently improved the attendance of urban youth? At what grade levels are these strategies most effective? And how do these strategies differ from other, less successful strategies?

Research should not just continue to cite the educational buzz words and emphasize the latest fads, like local school empowerment, schools of choice, promotion retention standards, etc.--but thoroughly indicate the pros and cons of these policy changes and successful ways of avoiding the negative implications.

Of course we know that research has been conducted to some extent in all of these areas. But in many cases, the research has not been comprehensive or explicit enough. The evidence of success is not measurable, or if it is measurable it does not show sustained effects. The research does not always clearly state how successful practices differ from what is being tried in many other school districts and the consequences for part.

And finally, the dissemination of these and other promising practices should be improved. Good educational strategies must not be determined by how well a sales representative can sell his/her product.

It is clear that the traditional methods are not working, and that many of us are not receiving objective information about promising practices. Just as representatives from various publishing companies "fight" to make the policy makers understand that their product is the best, methods need to be developed (other than sending it out in writing) to support the sharing of objective results with school board members, superintendents, curriculum administrators, etc. by knowledgeable persons who do not have a stake in the outcomes. Perhaps this should be done at the annual meeting of these various groups...Perhaps seminars should be held and paid for specifically for this purpose...Something must be done.

In summary, the research that is reported should clearly delineate the program or strategies that have been successful in working with urban youth. The results must be stated in measurable terms and show evidence that the effects have been sustained over time. The research should indicate comparison studies, why these programs/strategies work. And methods should be identified to disseminate successful practices, by objective spokespersons, to policy makers.

Gary Williamson
Director
Greensboro, N.C. Public Schools

There has been something of a revolution in the last few years in the development of statistical methods for the study of growth. There is a new framework for studies of individual growth, institutional aggregations of individual growth, and relations of other variables to academic growth. Formulations are in terms of individual mathematical models for growth and multilevel models for institutional (or other) effects.

A handful of researchers have brought about this new perspective. They include David Rogosa at Stanford University, John Willett at Harvard University, Tony Bryk at the University of Chicago, Steve Raudenbush at Michigan State University, Harvey Goldstein at the University of London, and others.

The appearance of this work in publication has been a phenomenon of the 1980's. The techniques have been appearing in print for more than half a dozen years, but applications are not yet commonplace. One would have thought that the educational community would have rushed to embrace a sounder, more unified approach to the description and investigation of academic growth.

Why have implementations of these methods not been widespread? I believe there are several reasons.

First, the theory has not been disseminated widely among practitioners. The work appears in a few journals, and in presentations at annual meetings of AERA and similar research organizations.

Second, there are few of the techniques in software accessible to the practitioner. AERA has offered summer institutes in the use of hierarchical linear model, and Bryk and Raudenbush have offered training in the use of their HLM program through this avenue. Other implementations are largely in the university research setting.

Another reason for the lack of applications lies in the nature of academic assessment. For longitudinal purposes, identical constructs must be measured on each occasion and measurement scales must be vertically linked; and, if comparisons of growth across different subject areas are of interest, horizontal equating is also mandatory. Commercially available tests may satisfy these characteristics to varying degrees. However, school systems usually test annually, so it takes a long time to build a data base. Teacher-made tests are administered frequently enough that a data base could be built quickly, but the psychometric characteristics of such tests are usually not known, and scales are not vertically or horizontally equated.

Another reason for the lack of longitudinal applications has to do with the organizational structure of most school district MIS and R&D offices. Such offices may not exist in smaller school districts. Even where they do exist, the groundwork for implementing and maintaining a longitudinal data base is extensive and involves the areas of measurement, data processing, and statistical analysis. Usually these functions are separated into different offices and the respective offices have their own charters with little overlap.

Where there is a systematic program of test management, it is often the case that one office (e.g., R&D) maintains the test data, and another office (e.g., MIS) maintains demographic data. Unless the organization makes it possible to pool the two sources of data (e.g., by scheduling time, personnel, and computer resources), it will be impossible to investigate the relationship between academic growth and other (institutional or individual) characteristics.

Another problem is the way that data are maintained. Even though many school districts have been using standardized tests for years, data files are maintained separately by year. Often data from different years vary in content, psychometric characteristics, format, storage medium, and perhaps even in geographic location (e.g., whether they are kept by the school, central office, etc.). Unless specific thought was given at the outset to setting up a longitudinal data base, or merging test data with other sources of student information, it is likely that no unique identification system is in place to allow linking student data from different sources. Although theoretically possible to create longitudinal data bases in such cases, the obstacles are considerable, and sometimes insurmountable.

Given the widespread interest in learning and accountability, this situation deserves further effort. I suggest that CRESST and other organizations consider the following issues and actions. Steps forward in these areas should help us all to make better efforts to assess student academic growth.

Issues for Research

- What are the existing practices of school districts with respect to the use of longitudinal data for the study of (1) individual student growth and (2) systematic individual differences in growth?
- What organizational characteristics determine the ability of a school system to establish and maintain a data base ideally suited for monitoring student growth (as opposed to status)?
- Do practitioners use data longitudinally? If so, how?
- What are optimal psychometric (and other) characteristics of tests that are designed to measure growth?
- What indicators of institutional quality most relate to individual academic growth?
- What is normative growth?

Areas for Development/Dissemination

- Enhance the ability of school systems to establish and maintain longitudinal data bases, by developing practical models that school systems may implement.
- Disseminate theoretical models for longitudinal data analysis.
- Develop commercial software implementations of recent methods for the analysis of longitudinal data.
- Seek ways to improve the longitudinal utilization of test results.

Lynn Winters
Palos Verdes Unified Schools

Issues Related to the School Reform Movement

One major concern of the local education agency is the effect of state curriculum and testing mandates on both the quality of instruction at the local level and the ability of the local education agency(LEA) to make its own decisions about curriculum. While states purport to provide leadership in educational reform which will lead to an improvement in education at the local level, the "school excellence" and "reform" movements have largely been implemented through assessment. In states such as California, the educational reform and excellence programs have been driven by testing programs followed by curricular documents which may or may not relate to the state tests. This model of school reform raises several issues related to testing policy:

1 What actual curricular changes occur after these "reforms-based" testing systems are in place?

2 How do state mandates for testing or curriculum reform affect options for local decision making and meaningful site based planning?

What instructional practices actually affect student learning and how are these practices related to the state-mandated curriculum and testing programs?

Issues Related to the Performance of "At-Risk" Students

A second concern of local education agencies is the performance of students euphemistically called "at-risk", the poor, the non-English speaking, and low performing minority students. There is much evidence that Black and Hispanic students do not perform as well on multiple choice tests (NAEP, SAT, CAP) as Asian and white students. What we have found, however, is an interesting phenomenon: Although Black students in our district do not perform as well on the multiple choice portions of the state test as white and Asian students with similar socioeconomic backgrounds, they do score as well on essay portions of the test.

The issues raised by this phenomenon are:

Is the "gap" between white and Black students in test performance an artifact of the testing process rather than a real difference in performance?

Or, as we move into performance testing, will we observe performance differences among ethnic groups with the same educational opportunities and socioeconomic status?

Assessing the Real Outcomes of Schooling

A third possible concern for a national testing center is the problem of assessing the real outcomes of schooling: organizational, thinking and problem solving skills. Currently these skills are best demonstrated in writing assignments, the development of projects (social studies, math, science) and in speaking tournaments. Issues associated with the assessment of these important outcomes include:

How are performance tests best developed and scored?

What constitutes a "portfolio" of work and how is this best evaluated?

How can performance tests be reported so that the public understands their results and will assign them the credibility currently given to norm referenced tests?

Assessing the Impact of Norm Referenced Reporting

A final issue of concern to LEA's that deserves notice is the affect of norm-referenced reporting systems on a systems ability to improve its instructional program. Current state testing programs essentially rank schools either on test performance or some other variable of "school effectiveness reports affect local curriculum? instruction? Supposedly these systems motivate schools to do better and provide information for improvement - is this indeed the case?

Further, I'd like to see better test reporting. I'd like to see us move away from using scores that have no direct interpretation, such as item response theory scores, and toward the use of content scales and descriptive information exemplified by the teacher reports of the grade 8 CAP writing assessment.

Thomas Kerins
Illinois State Board of Education

As the state director of testing in Illinois, I have been asked to provide testimony on the most pressing testing research and policy issues which should be addressed in future lab work. I will first comment on the assistance supplied by CRESST (Center for Research on Evaluation, Standards and Student Testing) over the last few years and then on future needs.

First, the Center, in cooperation with NCME, provided an admirable service by bringing together state testing directors, testing contractors, and university consultants to discuss mutual concerns and questions regarding the state RFP (Request for Proposal) process. The opportunity for these participants to openly discuss their problems and successes in the RFP process not only was a positive learning experience, but it immeasurably improved the quality of the RFP produced her by Illinois for its student assessment program. The precedent established here should continue through work on the Improving Large-Scale Assessment notebook.

Second, consulting assistance and encouragement by Dr. Robert Linn (Co-Director of CRESST) regarding the scoring and analysis procedures of the Illinois reading assessment enabled us to successfully establish the program. This assistance fits the ideal of the mission of CRESST--an organization that does not just regurgitate the easy answers, but presses forward on the difficult questions, an organization that is not afraid to tackle new positions.

A related initiative is the joint effort by CRESST and the Illinois State Board of Education in looking at a scoring system which would evaluate a student's writing ability and knowledge in a particular academic area, such as social studies. Time is such a precious resource; the more valid and reliable information we can efficiently squeeze from the data we collect, the more saleable and useful our assessment programs will become. Also, using model approaches in the state assessment programs can establish a precedent with regard to assessment practices for local schools to consider.

Fourth in the list of positive experiences was the opportunity to discuss testing issues with legislative staff from across the country. The earlier RFP meetings had documented the dilemmas that occur for state and local testing directors and contractors when communication problems lead to laws that are impossible for all of us, including CRESST personnel, to make workable.

All of these experiences have impressed me as positive actions that a national center should not only continue but expand.

Future Activities

Curriculum personnel and teachers are urging state assessment staff to develop methodologies for conducting large-scale assessments (thousands of schools, hundreds of thousands of pupils) that are more relevant for their need for information. They want to know how well pupils can "do science," not just respond to a single-answer multiple choice question. While the latter response is more efficient at the moment, the National Center could work to provide testing personnel with procedures that would enable them to efficiently assess pupils' knowledge and ability in ways that will make the information more useful.

The proliferation of mandated state and local assessments is producing a situation in which the goals of school improvement cannot be met because school personnel have neither the time nor the ability to turn assessment data into useful information for analysis or evaluation. Without the ability to do this, the tests become ends in themselves, rather than a means for school improvement. The Center needs to utilize emerging technical knowledge to develop ways that educators can better store, retrieve, and use assessment information with other indicators to evaluate students, curriculum, teaching, and administration.

Finally, monies should be included in the funding of the National Center to utilize staff expertise in the analysis of international assessments. We cannot as a nation recommend more sophisticated designs for the construction of international assessments without banking the monies needed to do the national analysis. A portion of the funds for a National Center for Research on Evaluation, Standards and Student Testing should be allocated to either conduct or manage the analysis responsibilities for the United States in international assessment.

George Malo
Tennessee State Department of Education

The perspective that I am taking comes from two areas, one from my current role in working with teacher evaluation in Tennessee and what I see as needs at the school-based level, and the second from talking with people in assessment at the state level.

I classify assessment issues at the school based level into three areas:

1. The teacher and the public understanding of assessment and its interpretation. More and more as we go out to evaluate teachers, we find them focused on only one type of assessment. They tend to look solely at paper and pencil measures. They are not familiar with many alternative assessment processes or the interpretation of such. I would like to see work done to help teachers become more familiar with the interpretation of different types of testing so that they know how to report student status and understand where a student stands in relation to content and the other parts of the curriculum in which they're teaching. Rather than just a broad range of where a student will stand, they need to characterize the complexity of student performance.

2. Approaches to assessment. We continually see that there needs to be more attention to multiple assessment of students. We need to consider assessments which better characterize student performance and better take into account student characteristics, that the teacher can consider in both assessment and teaching. For example, what are different types of assessment for different learning styles, what types of things do you look for in students, and what are the best approaches to testing at that level. Two issues seem salient here:

   a) With regard to content itself, what are some alternative ways of assessing content, looking at the same concepts or principles from a different light so that you can help students in terms of application.

   b) How do you go about teaching etiology--in other words, how do we help teachers to assess their own teaching strategies as a separate issue from assessing student learning.

3. The correlation between existing measures and new standards being set by the various professional groups. Recently, for example, the National Council of Teachers of Math have just come out with some of the standards which students should attain in mathematics at the various grades, as well as how to evaluate mathematics itself. Their emphasis is on more realistic and experience-based types of assessments rather than the old "one train goes south and another one goes north, which way does the wind blow?" or something like that, as I usually see in mathematics. What we're talking about is more practical, meaningful, experienced-based measures. How can you go about addressing this need, and how can we get more of a handle on critical thinking skills for which these new standards are looking.

John Keene has already talked about some of the state level issues that I see as crucial. We're looking at ways to customize our standardized testing so that in turn we can test some of those things that we think are particularly important for us as a state. We're also looking at ways to increase the efficiency of our testing. It seems that in Spring, three or four weeks are set aside for testing, and this seems to be taking away geometrically from academic learning time. For example, when we go to evaluate teachers they tell us not to send the evaluators out during the whole month of March because all they are doing is testing. So we are looking at ways to "kill two birds with one stone" and not be testing every day.

A concern is how the state can help teachers more with diagnostic types of testing that are useful in their instruction, rather than solely standardized achievement testing.

Edward D. Roeber
Michigan Department of Education

It is a pleasure to take this opportunity to comment on the activities of the UCLA Center for Research on Evaluation, Standards and Student Testing (CRESST). I have had the pleasure of working with the Center in several different ways over the past few years and I am able to comment on how these activities have been useful to the Michigan Department of Education and our local school districts. My work with CRESST began as they initiated the development of their initial proposal for funding and continues to today.

The first thing that struck me as I worked with Eva Baker a few years ago when she was first putting together a proposal for CRESST was her willingness to work with state testing people to identify research needs and priorities. Previous testing centers had ignored testing going on at the state level, in spite of the fact that most states have some type of state testing, state assessment, or student competency testing program. Over the life of the Center, state testing activity has increased immensely, since student testing is one way that policy makers have used to determine the need for changing the educational system, for bringing about those changes, and monitoring the impacts of those changes. State testing represents a significant proportion of student testing that occurs in the United States, so it should also occupy a significant research priority for a research center on student testing. I have been pleased with what I have seen as CRESST's response to this priority.

One way that CRESST has been trying to improve the quality of student assessment at the state level has been through a careful examination of the manner in which test development and test administration contractors are procured. A task force representing state testing directors, measurement specialists and testing contractors has examined this issue at some length and has developed guidelines for states to use in developing requests for proposals and selecting contractors. These represent both practical yet technically correct advice for state assessment staff and others to use.

Another way in which CRESST has (and is) attempting to have a positive effect on student testing has been through reaching out to groups such as the National Conference of State Legislatures and the National Governors Association. By making presentations on student testing to these groups, the Center is beginning the process of establishing a dialogue with the policy-making groups that often establish the parameters for student testing at the state level. CRESST has begun the important process of educating these groups on both the technical and practical constraints on student testing at the state level. Such efforts need to continue and to be expanded. Policy makers have real educational concerns and view testing as a means to address these. Whether testing is used to monitor education, to change it, or to evaluate changes made to it, testing needs to be carried out in a technically and educationally sound manner. CRESST should continue to try to influence the groups and individuals who establish testing programs at the national, state, and local levels.

A third way in which the Center can and is having an impact on states is by identifying areas for research on new assessment strategies and practices. For example, in the area of reading, several states are exploring alternative means of assessment. Both Michigan and Illinois have worked with Center staff on this. Having a Center such as CRESST available to help us is important for two reasons. First, some of the most innovative work in developing better ways of assessing students is occurring now at the state level–in states such as Michigan, Illinois, California and Connecticut. Second, although much work is going on at the state level, states rarely have the opportunity (or funding) to carry out the basic research needed in these areas. An emerging area of concern for us in Michigan is designing better means of assessing the areas of science and social studies. CRESST can do even more by linking states with the university researchers and helping the network of persons to carry out basic and applied research needed by the states to support fundamental changes in assessment practices.

For these reasons and more, I do hope that CRESST's funding will be continued. There are important challenges which lie ahead. I do hope that the center for student testing will continue to work collaboratively with states and I hope that the formal and informal mechanisms for doing this can be enhanced. Currently, some of the collaboration occurs only because states recognize the importance of doing so and are willing to contribute staff time and travel funds to do this. This limits the effort to states which have the available resources. I hope that more formal mechanisms for relating CRESST to the research priorities and needs of the states can be built into the grant application process. Recognizing that some of the best work in innovative assessment practices is occurring at the state level, CRESST should be funded to establish and maintain a state-based research effort. What innovations are occurring? How can CRESST assist states? By working together with them on common research areas and topics. One way to do this would be by formalizing a network of state testing programs; another would be by setting aside a certain proportion of the funds for research and development activities directly related to statewide student programs. The research and development activities could then be directed by the state network and CRESST. It is my hope that CRESST researchers would work directly with state testing staff and their consultants in carrying out research of mutual interest.

As I indicated at the outset, it has been a pleasure to work with the CRESST staff and faculty on the technical and the practical problems of student testing at the state level. I hope that the important work of CRESST will be continued, so that the work that we have developed collaboratively over the past few years can be continued and used as a basis for improving the development and use of tests. With the emphasis now being given to student testing at the state level, it is important that the resource which has been developed in CRESST be continued and enhanced.

John Keene
National Computer Systems

Faced with this task of identifying research and development priorities, I thought I'd just write down some questions. Then I started organizing them. I put down 18 questions and organized them in 14 areas, starting from more general to specific (see Appendix).

My first area is demystifying test scores. I work for a test publisher and I build NRT tests. I find that there is much too much emphasis placed on the tests. The over-emphasis on the tests is especially in two areas: the political arena, where the legislatures are mandating the tests, and secondly with the general public. There are some questions I'd like answered. What are the effects of test use from legislative mandates? In other words, does mandating a certain test for grades 4-8 help the educational process? What do we know about that? We see test scores going up, but is that really a valid criterion? What's the outside validation? Further, let's learn from positive examples. Where and how does state mandated testing improve the educational process in general? Is this a good model to be using?

With the general public, the tests are also mystifying. I heard today on the news that some researchers in Georgia had found that the Verbal SAT scores do not correlate well with reading performance in college. It was big news, splashed all over the headlines. I know that since that test is mostly vocabulary, it may not correlate well with reading. The public, however, doesn't understand all that. We ought to find out what the public does know about assessment, and then try to focus our reports from that perspective. Also, how can we educate the public about the proper use of tests and test scores? As an NRT publisher, we have a million phone calls a day from people asking, "What does this mean?" "What good is this?" "How can I use this test?" I think we need to know how to educate the public. Right now I don't believe that they know much about tests in general.

Then, once we demystify testing in the eyes of the public and the politicals, we've got the educators. I would like to know what the educators think about the effects of tests on instruction. What is the effect of teaching to the test on the educational process? We've been holding results from mandated tests, saying "Go after this, these are the things you should teach." How is that affecting instruction? My second question is, how can measurement-driven instruction yield better results? If we're going to use such a model, how can it work better? Also with the educators, we need to find out what they think are the proper, practical, and feasible uses of tests. My question here is, can we describe existing models of proper and effective test use? If we can, what do teachers think about the ideal use of tests in their present use?

I move on now to a more practical area, that of managing research, which is something we have to do when we build an NRT. One of our biggest problems is getting school districts to cooperate with us in getting norms. How can we obtain more student data for norm development? We're having a horrible time, and any NRT publisher will tell you that. How can we work it out with the schools so we don't take up so much of their time, but yet we get the norms that we need? If we could answer that question, we would save a lot of money. Second, with regard to the annual norms issue, what types of models for annual norms are appropriate? User data, user plus non-user data, whole new standardization, what makes the most sense? How are these different; how do we distinguish the differences when we look at different publishers' products?

On the subject of implementing research--I don't know any place to go for help. I know that one good book I have always used is *Educational Measurement*, and the new edition is just out. But even beyond that there isn't a place where we defined our knowledge about testing, evaluation, and instruction. We need some type of anthology that tells what we know about instruction and test assessment, and so on. Further, we need to take all this knowledge and research we have done and report at meetings such as AERA and figure out how we can transform it into practical applications. That's another large question.

On the second page of my attachment are questions about what I spend a lot of my time doing. These are harder research questions but I think that they deserve attention. A lot of states and districts are using what they call derived norms or equated norms, meaning they're not really giving the NRT test but they're giving a part of the NRT test or a test that's been equated to it. One of the big questions is: are these equated or derived norms valid for Chapter 1 testing? We need real research on that. Another, broader question is, what restriction should be placed on the reporting of derived or equated norms? Should we pretend they are the actual norms, or should some sort of qualifiers be required?

Another related area is content alignment. Everybody wants their tests to test their content, and not the content on your standardized tests. So can we look into content hierarchies, and can we develop one so that all test items and instructional objectives can be classified? I guess that's an ideal, but I'd love to see it. It would help us out. Also, how meaningful are the small content differences between two curricula or two tests? I don't know the answer to that, but we're hoping to look into it. A third is--we have NRT scales, and we compare students from one to the other on the basis of a percentile on that test. It would really be nice to have a valid and reliable content scale, where I can look at a student and describe their position in a content hierarchy or dimension. NAEP has attempted some of this with their reading scales, but we really need even better descriptions. Tying the content alignment together, is it possible to develop content reference scales to replace norm reference comparisons? Lastly, what are the bounds of the vertical scaling you can do with the content in basic subjects? Thank you.

James Olsen
WICAT Associates

My primary areas of interest in improving testing and evaluation cluster under the title, "we ought to look at learning improvement rather than just achievement." And by that, I mean that we must examine all the aspects of the learning process, going beyond just the material that we can evaluate on traditional tests.

I think there's a need for initiating evaluations of advanced technology applications for schools. We're seeing a large number of systems--microcomputers, integrated learning systems, interactive video systems--that are being implemented in schools with the intent that they're going to benefit students, but there's been little carefully controlled research on the process and outcomes of those systems. Nor has there been sufficient attention concerning how to prepare teachers to deal with those types of technologies, which are different from what they've used in the past.

I think there is a need for development, implementation and evaluation of authentic performance-based assessment models. We are starting to see some excellent prototype models being attempted, and such attempts ought to be encouraged and transferred to new content domains.

I think we ought to be investigating innovative types of assessment that will be needed for the year 2001. We should be thinking ahead to what the twenty-first century will be like, particularly concerning future careers that our students will be entering, careers for which we should now be trying to prepare our students. Even though we ourselves may not know what the world is going to be like in 2001, we need to maintain a many-year-ahead vision that will help us to look far down the road rather than just ahead.

I think there's an important need for the articulation of objectives and content-referenced scales across grade levels in criterion-referenced tests. Such instruments would enable us to trace the performance and progress path of a student all the way through his or her school career, providing invaluable information to the teacher. At the beginning of a new school year, as the classes are being restructured and as new students enter, teachers could know the scale structure, achievement and curriculum scale positions for students coming into their classes.

Furthermore, such a system would fulfill the strong need for increased longitudinal research on learning improvement. We need longitudinal research to examine a variety of current policy issues. For example, we need research on at-risk and drop-out prevention programs. We need to assess the effects of year-round schools, particularly in light of some of the budget cuts that schools are facing now. We need to examine and evaluate the effects of early childhood education programs. Similarly, there are recent, urgent calls for the radical restructuring of our educational systems and schools. What will be their effects? And what of court-ordered takeovers of schools if they're not performing according to stated criteria? How will these outcomes be reliably and validly identified?

Finally, growing from our rapidly expanding global communication networks, there are interesting opportunities for collaborative learning among students, schools, districts, states, the nation, and the world. There's a project that is dealing with satellite data collection on important science problems where high school students are now becoming practicing scientists. These students have access to more data on some critical science problems than yesterday's scientists. If we can engage high school students as scientists, and/or in collaborative global projects like this one, I think we'll make sure that our schools are more interesting and informative places to be. In such an environment, we will be able to keep students motivated in school and not have to worry about trying to recover them or about trying to prevent dropouts. I'd like to see schools which are seen as exciting places for students, where students are well prepared for future careers, and particularly for careers in which they want to spend the rest of their lives.

Stephanie Burton
National Computer Systems

Many of the testimonies today strongly recommend that we focus research dollars on innovative programs of a more practical nature. In other words, there seems to be a good deal of frustration with the use and misuse of standardized tests and a good deal of interest in piloting new ways of evaluating students as individuals in order to return the focus of education to helping the kids instead of winning the elections. While education certainly won't change overnight, innovative means of individuals in order to return the focus of education will be feasible through advancements in technology. However, it seems clear that measurement of the efficacy of that innovation within and across populations will remain a requirement for American education to gauge and report its progress. In that light, perhaps the most significant research contribution to education in the 1990s would be the development of a "common language" to be used across the educational spectrum in order to support future innovation in measurement and technology.

Student mobility within and across district and state boundaries continues to accelerate. This increased mobility is changing the cultural and socio-economic make-up--and thus the educational needs--of students, classrooms, schools, and school districts--almost overnight. I can think of no greater contribution research could make to the future of eduction than to provide a common language to describe educational achievement and learning needs across geographic boundaries.

Instead of creating common curricula, test items and performance measures, development of an "educational Esperanto" might provide a standard taxonomy for describing educational experience, objectives, test items, and performance measures as well as common definitions for SES descriptors, demographics for individuals, schools and districts, entitlements, disabilities, etc. Instead of standardizing outcomes, accepted common definitions of data elements might facilitate and support innovation in methods of measurement, customized to the distinct and changing needs of the population both within and across traditional boundaries. Through a common language framework, technology could support an "electronic transcript," moving with the individual throughout his or her educational life and providing data elements which enable each educational institution to describe, in its own terms, the educational needs and abilities of that individual.

I look forward to the follow-up on the testimony given at AERA. It is clear that there is no dearth of educational subject matter pleading for research. This means the challenge will continue for those of us whose interest is in education.

David Berliner
Arizona State University

It is clear that we all recognize that standardized testing is desired by legislators and the public. It is equally obvious that such testing has a negative effect on instruction, in that it trivializes a good deal of the curriculum, since it drives the curriculum. Because the tests drive the curriculum, it is important to develop tests of the things in which we are more interested: tests of attitudes toward science, tests of understanding of the process of mathematics and science, tests of literary criticism, etc. We need tests that measure the kinds of outcomes we hope for in English, History, Science, and Mathematics, but about which we do nothing during instruction. We instruct, typically, in those things that are on the current standardized tests. Those are usually low level skills, or facts, names, dates, etc. Good tests of process and attitudes are needed. Perhaps then they can drive the curriculum. Some of that work is going on. I know that National Science Foundation has a project like that. But we need tests like these in other intent areas as well.

In addition, it is clear that student portfolios and projects, and teacher portfolios that document their contributions, are important ways of presenting evidence of accomplishment in schooling or teaching. We have no good measurement procedures for comparing portfolios or for grading projects. I think there is a considerable literature on scaling that could be applied to rating the vastly different portfolios and projects presented by students and the very different portfolios that are presented by teachers who might want advancement in career ladders. (I see the two problems as the same--each requires finding ways to scale very different phenomena.) It strikes me that a contribution in this area is needed. Otherwise, we keep having those sensible records of accomplishment considered to be unmeasurable. I think a vigorous evaluation of those materials is possible. I hope these comments help.

Merlin C. Wittrock
University of California, Los Angeles

Process-Oriented Tests and Teaching

Tests influence teaching. Tests of comprehension, tests of student preconceptions, and tests of the strategies and the cognitive skills students use to learn in school can significantly contribute to improving teaching in America.

From a wide variety of recent research studies on teaching (cf. *The Handbook of Research on Technology,* M.C. Wittrock, 1986), it is clear that student background knowledge, learning strategies, and metacognitive processes play critical roles in influencing the learning of subjects taught in schools. Recent research on preconceptions in science, on addition and subtraction strategies in mathematics, on comprehension in reading, on attention in learning disabilities, and on attribution processes in motivation show impressive gains (e.g., 100% to 200% in reading comprehension) in learning when teaching builds upon information about students' background knowledge and thought processes.

We have the research base needed for making sizable gains in student learning. To realize this research-based potential we need to construct practical and useful process-oriented tests that teachers can use to measure student preconceptions, comprehension processes, and thinking skills.

These new tests will provide diagnostically and in instructionally useful data directly applicable to the design of instruction for individual students, including at-risk students. These tests will also facilitate the evaluation of student comprehension of fundamental subject matter concepts taught in schools, in areas such as social studies, history, science, mathematics, and reading.

Our nation faces serious challenges in the improvement of teaching in its schools. In many of our nation's schools we must emphasize student comprehension and student understanding better than we have in the past.

We now have the knowledge needed to improve the teaching and the testing of comprehension. We need to develop a new approach to testing, in which comprehension and understanding are the focus of evaluation, and student preconceptions and learning processes are the focus of measurement.

These process-oriented tests will not replace standardized achievement tests. Instead, they will serve different, much-needed functions, including the important one of focusing teaching on understanding students' background knowledge, beliefs, attributions, and thought processes, and on the attainment of student comprehension.

Because tests influence teaching, tests of student understanding and of student thought processes will enhance the teaching of comprehension and the learning of meaningful and lasting knowledge.

Appendix

ASSESSMENT ISSUES - RESEARCH QUESTIONS

I. DEMYSTIFYING TEST SCORES

A. POLITICAL INFLUENCE

1. WHAT ARE THE EFFECTS ON TEST USE FROM LEGISLATIVE MANDATES?

2. WHERE AND HOW DOES STATE MANDATED TESTING IMPROVE THE EDUCATION PROCESS?

B. GENERAL PUBLIC

1. WHAT DOES THE PUBLIC UNDERSTAND ABOUT ASSESSMENT?

2. HOW CAN WE EDUCATE THE PUBLIC ABOUT THE PROPER USES OF TESTS AND TEST SCORES?

II. EDUCATING EDUCATORS

A. INSTRUCTIONAL FOCUS

1. WHAT IS THE EFFECT OF TEACHING TO THE TEST ON THE EDUCATIONAL PROCESS?

2. HOW CAN MEASUREMENT DRIVEN INSTRUCTION YIELD BETTER RESULTS?

B. PROPER USE OF TESTS

1. CAN WE DESCRIBE EXISTING MODELS OF PROPER AND EFFECTIVE TEST USE?

2. WHAT DO TEACHERS THINK ABOUT THE IDEAL USE OF TESTS AND THEIR PRESENT USE?

III. MANAGING RESEARCH

A. OBTAINING COOPERATION

1. HOW CAN WE OBTAIN MORE STUDENT TIME FOR NORMS DEVELOPMENT?

2. WHAT TYPES OF MODELS FOR "ANNUAL NORMS" ARE APPROPRIATE?

B. IMPLEMENTING RESEARCH

1. WHAT DO WE KNOW ABOUT TESTING AND INSTRUCTION?

2. HOW DO WE TRANSFER KNOWLEDGE AND RESEARCH FINDINGS AND APPLY THEM TO PRACTICAL CONCERNS?

IV. CUSTOMIZED TESTING

A. DERIVED NORMS

1. ARE EQUATED OR DERIVED NORMS VALID FOR CHAPTER TESTING?

2. WHAT RESTRICTIONS SHOULD BE PLACED ON THE REPORTING OF DERIVED OR EQUATED NORMS?

B.CONTENT ALIGNMENT

1. CAN A CONTENT HIERARCHY BE DEVELOPED SO THAT ALL TEST ITEMS AND INSTRUCTIONAL OBJECTIVES CAN BE CLASSIFIED?

2. HOW MEANINGFUL ARE SMALL CONTENT DIFFERENCES WHEN RELATED TO TEST PERFORMANCE?

C. CONTENT SCALES

1. IS IT POSSIBLE TO DEVELOP CONTENT REFERENCED SCALES TO REPLACE NORM-REFERENCED COMPARISONS?

2. WHAT ARE THE BOUNDS OF VERTICAL SCALING OF CONTENT FOR THE SUBJECTS OF READING, MATHEMATICS, AND LANGUAGE?