The CRESST Line
Winter 2002

Newsletter of the National Center for Research on Evaluation, Standards, and Student Testing

FROM THE DIRECTORS: Standards for Educational Accountability Systems
CRESST Co-Directors Eva L. Baker, Robert L. Linn, Joan L. Herman, and CRESST Associate Director Daniel Koretz

THE passage of the education reform law has spotlighted testing and accountability once again. Provisions to test students in Grades 3-8, to develop approaches for measuring adequate yearly progress, and to reach full proficiency in 12 years are among the salient features of the law that states will begin to address. While the details of implementation remain to be worked out, it is clear that all states will now review the present form of their testing programs and accountability systems to determine how they will be changed to meet these new expectations. Now is the time for states, as they reflect and prepare for action, to consider anew the true quality of their future efforts. What gauge should be used to determine the quality of accountability plans and operations?

We believe that research, development, and evaluation knowledge can assist states in sorting through their options and in improving quality. CRESST, in partnership with the Consortium for Policy Research in Education (CPRE), with the Education Commission of the States (ECS), and with advice and review from numerous colleagues in research and practice, offers the Standards for Educational Accountability Systems. These standards are intended to provide guidance to states and districts in conducting self-reviews of their own systems and to delineate criteria by which developing accountability systems can be judged. The Standards for Educational Accountability Systems represent compiled knowledge developed from sources including the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement on Education [AERA, APA, & NCME], 1999), research findings on testing and accountability systems, and studies of best practices. Many of their precepts are based on public access to the composition of the system and reporting of results.

Because experience with accountability systems is still developing, the standards we propose are intended to help evaluate existing systems and to guide
the design of improved procedures. The standards strongly endorse each state’s responsibility to conduct continuing evaluation of its own accountability system. It is not possible at this stage in the development of accountability systems to know in advance how every element of an accountability system will actually operate in practice or what effects it will produce. Evaluations, conducted in-house, or by universities, external organizations, or teams of experts, are essential if states are going to learn systematically from one another and for the nation to judge the effectiveness of their efforts for children. Evaluation results will be essential to the continuing improvement of testing programs and accountability provisions.

In sum, the standards offered below represent models of practice derived from three perspectives: research knowledge, practical experience, and ethical considerations. They should be conceived of as targets for state and local systems and as criteria to judge proposed models of accountability development.

It should be understood that tests included in an accountability system should meet the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999). What we have highlighted here are criteria that apply especially to accountability systems. It is likely also that additional standards will be subsequently developed based on evaluations of accountability system effects. A more comprehensive version of the standards, including comments, will be published as a forthcoming CRESST Policy Brief.


A. Standards on System Components

1. Accountability expectations should be made public and understandable to all participants in the system.

2. Accountability systems should employ different types of data from multiple sources.

3. Accountability systems should include data elements that allow for interpretations of student, institution, and administrative performance.

4. Accountability systems should include the performance of all students, including subgroups that historically have been difficult to assess.

5. The weighting of elements in the system, different test content, and different information sources should be made explicit.

6. Rules for determining adequate progress of schools and individuals should be developed to avoid erroneous judgments attributable to fluctuations of the student population or errors in measurement.
B. Testing Standards

7. Decisions about individual students should not be made on the basis of a single test.

8. Multiple test forms should be used when there are repeated administrations of an assessment.

9. The validity of measures that have been administered as part of an accountability system should be documented for the various purposes of the system.

10. If tests are to help improve system performance, there should be information provided to document that test results are modifiable by quality instruction and student effort.

11. If test data are used as a basis of rewards or sanctions, evidence of technical quality of the measures and error rates associated with misclassification of individuals or institutions should be published.

12. Evidence of test validity for students with different language backgrounds should be made publicly available.

13. Evidence of test validity for children with disabilities should be made publicly available.

14. If tests are claimed to measure content and performance standards, analyses should document the relationship between the items and specific standards or sets of standards.

C. Stakes

15. Stakes in accountability systems (or incentives and sanctions) should apply to adults and students and be coordinated to support system goals.

16. Appeal procedures should be available to contest rewards and sanctions.

17. Stakes for results and their phase-in schedule should be made explicit at the outset of the implementation of the system.

18. Accountability systems should begin with broad, diffuse stakes and move to specific consequences for individuals and institutions as the system aligns.

(see From the Directors, page 4)
D. Public Reporting Formats

19. System results should be made broadly available to the press, with sufficient time for reasonable analysis and with clear explanations of legitimate and potential illegitimate interpretations of results.

20. Reports to districts and schools should promote appropriate interpretations and use of results by including multiple indicators of performance, error estimates and performance by subgroup.

E. Evaluation

21. Longitudinal studies should be planned, implemented, and reported evaluating effects of the accountability program. Minimally, questions should determine the degree to which the system:
   a. builds capacity of staff;
   b. affects resource allocation;
   c. supports high-quality instruction;
   d. promotes student equity access to education;
   e. minimizes corruption;
   f. affects teacher quality, recruitment, and retention; and
   g. produces unanticipated outcomes.

22. The validity of test-based inferences should be subject to ongoing evaluation. In particular, evaluation should address:
   a. aggregate gains in performance over time; and
   b. impact on identifiable student and personnel groups.

New Updated Web Site for The National Research and Development Centers

A Web site providing links to research reports and publications for the 12 National Research and Development Centers sponsored by the U.S. Department of Education’s Office of Educational Research and Improvement is now fully operational with a powerful new search engine and updated content. The updated site location is: http://research.cse.ucla.edu
CRESST at the 2002 American Educational Research Association and National Council on Measurement in Education Meetings

The following is a list of CRESST partners and many of their sessions at the 2002 AERA/NCME meetings in New Orleans. We have tried to make the list as accurate as possible, but encourage you to check your program guides for complete schedules. Because of space limitations, the list reflects only CRESST partners, not all colleagues participating in the sessions. Our thanks to those colleagues as well. We have listed only the primary university affiliation for each researcher.

APRIL 1, MONDAY AFTERNOON

1.31 Division D New Member Poster Session—Measurement and Qualitative Research Methods
Noon-1:30 p.m. Sheraton, Pontchartrain C, 3rd floor
Aspects of Science Achievement and Their Links to Test Items
Presenter: Richard Shavelson, Stanford University

1.73 Teacher Evaluation and Performance Assessment: Reviews, Trends and Vignettes
Noon-1:30 p.m. Marriott, Studio 9, 2nd floor
Vignettes as an Alternative Teacher Evaluation Instrument: A Pilot Study
Presenter: Maria Ruiz-Primo, Stanford University

4.49 Tools of the Trade: Measuring Effectiveness of Reforms
2:15-3:45 p.m. Sheraton, Grand Ballroom E, 5th floor
Evaluating Creative Teaching Practice and Collaborative Learning Environment: Quantitative and Qualitative Indicators in Evaluating School Reform
Presenters: Ann Mastergeorge and Ingrid Roberson, UCLA

6.22 New Member Research in Motivation
4:05-5:35 p.m. Sheraton, Pontchartrain C, 3rd floor
Effects of Gender, Achievement Level, and Audience on Adolescent Students’ Communicated Attributions and Affect
Presenter: Jessica Howie, UCLA
The Impact of KidzLit on Student Attitudes Towards Reading
Presenters: Halle Aten, Jessica Howie, and Eva Chen, UCLA

6.37 From Large-Scale Assessments to Classroom Practice: Issues in Designing Assessment Systems That Support Learning
4:05-6:05 p.m. Sheraton, Grand Ballroom A, 5th floor
Building Bridges Between Classroom and Large-Scale Assessments
Presenter: Lorrie Shepard, University of Colorado at Boulder

APRIL 2, TUESDAY MORNING

11.39 Improving Educational Assessments: Protocol Analysis Methodologies for Addressing Validity Issues
8:15-9:45 a.m. Sheraton, Bayside B, 4th floor
Cognitive Validity of Performance Assessments via Protocol Analysis: Technical Aspects
Presenter: Richard Shavelson, Stanford University

11.53 Expanding Foundations of Evaluation Practice
8:15-10:15 a.m. Le Meridien, Orleans, 3rd floor
Oxymoronic Program Evaluation: The Short-Term Longitudinal Analysis Dilemma
Presenters: Pete Goldschmidt and Silvia Swigert, UCLA

APRIL 2, TUESDAY AFTERNOON

15.46 High-Stakes Testing and Accountability
12:25-1:55 p.m. Sheraton, Grand Ballroom E, 5th floor
Discussant: Brian Stecher, RAND

15.51 Valid Assessments: For Whom and for What?
12:25-1:55 p.m. Le Meridien, France III, 3rd floor
Discussant: Robert Mislevy, University of Maryland

15.69 Schooling Poor and Minority Adolescents: Reflections on Validity and Value in Education Research
12:25-1:55 p.m. Sheraton, Rhythms III, 2nd floor
The Possibilities and Limitations of School Reconstitution
Presenter: Heinrich Mintrop, UCLA

17.4 Evaluating Technology Impacts: Assessing Student Technology Outcomes
2:15-3:45 p.m. Marriott, La Galerie 1, 2nd floor
The Validity of a Sampling Approach for Assessing Student Performance Using Science Notebooks
Presenters: Maria Ruiz-Primo and Richard Shavelson, Stanford University
Discussant: Brian Stecher, RAND

19.38 The Science in Education Research: Implications of a National Research Council Study
4:05-6:05 p.m. Sheraton, Grand Ballroom E, 5th floor
Chair: Richard Shavelson, Stanford University

APRIL 3, WEDNESDAY MORNING

24.02 Schools, Classrooms and Students as Units of Analysis: New Evidence on the Components of Variance of Student Achievement and Social Status From Large Cross-National Studies
8:15-10:15 a.m. Marriott, La Galerie 3, 2nd floor
Exploring the Variance in Eighth-Grade Mathematics Performance in TIMSS
Presenter: Daniel Koretz, Harvard University
Discussant: Richard Shavelson, Stanford University

24.73 What Does It Mean to Learn in the Workplace? Differing Inter- national Perspectives
8:15-9:45 a.m. Sheraton, Rhythms III, 2nd floor
Discussant: Lauren Resnick, University of Pittsburgh

26.1 Teacher Learning and Educational Professional Development: Establishing Validity in Research
10:35 a.m.-12:05 p.m. Marriott, La Galerie 3, 2nd floor
Presenter: Hilda Borko, University of Colorado at Boulder

(see AERA, page 6)
26.62 Prepared for Life: Student Competencies in International Comparison (OECD-PISA)
10:35 a.m.-12:05 p.m. Le Meridien, St. Jerome, 3rd floor
Discussant: Richard Shavelson, Stanford University

APRIL 3, WEDNESDAY AFTERNOON

28.38 Social Dimensions of Technology Environments
12:25-1:55 p.m. Sheraton, Ponchartrain E, 3rd floor
Social Cues in Multimedia Learning: Role of Speaker’s Voice and Conversational Style
Presenter: Richard E. Mayer, University of California, Santa Barbara

30.5 How Reform Ideas Travel
2:15-4:15 p.m. Marriott, La Galerie 1, 2nd floor
Discussant: Lauren Resnick, University of Pittsburgh

30.14 The Alignment of Standards and Assessment: Building Better Methodologies
2:15-4:15 p.m. Marriott, Mardi Gras D, 3rd floor
Alignment and College Admissions: The Match of Expectations, Assessments, and Educator Perspectives
Presenters: Joan L. Herman, Noreen Webb, and Stephen Zuniga, UCLA
Integrated Alignment: The LEARNOME as a Descriptive Tool to Map Functional Relationships
Presenter: Eva L. Baker, UCLA

30.40 The Design and Delivery of Distance and Distributed Instruction
2:15-4:15 p.m. Sheraton, Gallier, 4th floor
Validating Link Categories for Automated Classification of Concept Map Links

30.45 The Impact of Technology on Assessment: New Opportunities for Knowing What Students Know
2:15-4:15 p.m. Sheraton, Grand Ballroom C, 5th floor
Inexorable and Inevitable: The Continuing Story of Technology and Assessment
Presenter: Randy Bennett, Educational Testing Service
The Roles of Technology in the Assessment Argument
Presenter: Robert Mislevy, University of Maryland
Discussant: Robert L. Linn, University of Colorado at Boulder

30.47 Transitions in Teaching
2:15-3:45 p.m. Sheraton, Grand Ballroom E, 5th floor
Transitions in Teaching: An Examination Across Cases
Presenter: Hilda Borko, University of Colorado at Boulder

30.55 Validity and Value of Education Research for Parents: What Parents Need to Know
2:15-3:45 p.m. Le Meridien, Rosalie, 3rd floor
Presenter: Ronald Dietel, UCLA

30.66 Comparing Hands-On Inquiry to Textbook-Based Science in Elementary Classrooms
2:15-3:45 p.m. Sheraton, Rhythms III, 2nd floor
The Value of HLM in Conducting Analyses of Student Science Learning
Presenter: Michael Seltzer, UCLA
Discussant: Richard Shavelson, Stanford University

APRIL 4, THURSDAY MORNING

37.14 Innovative Measurement of Students’ Scientific Understanding
8:15-9:45 a.m. Marriott, Mardi Gras H, 3rd floor
Problem Solving Strategies in Performance Assessment: Hands on and Minds on
Presenter, Richard Shavelson, Stanford University

37.31 SIG-Computer Applications in Education Poster Sessions
8:15-8:55 a.m. Sheraton, Ponchartrain C, 3rd floor
Discussant: Derek Mitchell, UCLA

37.46 Psychosocial and Political Factors: Policy Effects on Teacher Motivation, Parent Trust and Student Identity
8:15-9:45 a.m. Sheraton, Rampart, 5th floor
Performance Motivation in Schools on Probation in Maryland and Kentucky
Presenter: Heinrich Mintrop, UCLA

39.50 Advances in Growth Modeling: Taking Into Consideration Where Students Start in Studying How Rapidly They Progress
10:35 a.m.-12:05 p.m. Le Meridien, France I, 3rd floor
Examining Relationships Between Where Students Start and How Rapidly They Progress: Some Implications for Monitoring School Performance
Presenters: Michael Seltzer, Kilchan Choi, and Yeow Meng Thum, UCLA
Modeling Heterogeneity in Relationships Between Initial Status and Rates of Change: Latent Variable Regressions in Three-Level Hierarchical Models
Presenters: Kilchan Choi and Michael Seltzer, UCLA
Early Identification of Reading Problems Using Growth Mixture Models
Presenters: Christy Boscardin and Bengt Muthén, UCLA
Modeling Heterogeneity of Growth Trajectories Depending on Initial Status
Presenter: Andreas Klein, UCLA
Measuring Progress Toward a Goal: Inference for the Ratio of Change
Presenter: Yeow Meng Thum, UCLA

39.63 Issues Concerning Classification of English Language Learners
10:35 a.m.-12:05 p.m. Marriott, Bonaparte, 3rd floor
Issues and Problems in Classification of ELL Students: Analyses of Extern Data
Presenter: Jamal Abedi, UCLA

APRIL 4, THURSDAY AFTERNOON

41.59 On Reliability, Validity, Evaluation and Instruction: A Symposium in Memory of Lee J. Cronbach
12:25-1:55 p.m. Sheraton, Grand Ballroom A, 5th floor
Chair: Richard Shavelson, Stanford University
Validity Theory
Presenter: Robert L. Linn, University of Colorado at Boulder
Aptitudes and Instruction
Presenter: Haggai Kupermintz, University of Colorado at Boulder

43.27 Structured Poster Session
2:15-3:45 p.m. Sheraton, Rhythms II, 2nd floor
Online Training to Evaluate Student Work
Presenters: David Niemi and William Bewley, UCLA
Computer Applications in Education Paper Discussions
2:15-2:55 p.m. Sheraton, Armstrong Ballroom, 8th floor
Implementation of a Problem Solving Software in High School Science Classrooms
Presenters: Eva Chen, Bruce Burnam, Halle Aten, Veena Nambiar, and Jessica Howie, UCLA

Examining the President’s Annual Testing Agenda: Challenges for the Educational Community
4:05-6:05 p.m. Mariott, La Galerie 4, 2nd floor
Visions of Test Results Dance in Their Heads
Presenter: Eva L. Baker, UCLA
Technical Considerations in the Use of NAEP to Confirm States’ Achievement Gains
Presenter: Edward Haertel, Stanford University
Annual Testing Requirements: Some Tradeoffs
Presenter: Robert L. Linn, University of Colorado at Boulder

Learning in Virtual Reality Environments
8:45-9:25 a.m. Sheraton, Armstrong Ballroom, 6th floor
The Role of Immersion in Virtual Reality Environments: How Does Media Affect Students’ Learning?
Presenter: Richard E. Mayer, University of California, Santa Barbara

Thinking, Problem Solving and Argumentation in Technology Environments
8:15-10:15 a.m. Sheraton, Gallier, 4th floor
Computer-Based Assessment With IMMEX: Linking Cognitive and Online Problem Solving Processes

Item Response Theory as Non-Linear Mixed Models: Implications for Modeling Items and Persons by Substantive Features
8:15-9:45 a.m. Sheraton, Grand Ballroom B, 5th floor
Discussant: Brian Junker, Carnegie Mellon University

Applications of HLM in Studies of School Effects
10:35 a.m.-12:05 p.m. Sheraton, Rhythms I, 2nd floor
Discussant: Michael Seltzer, UCLA

Research on Learning and Teaching
10:35-11:15 a.m. Sheraton, Pontchartrain C, 3rd floor
Cognitive Interpretation of Performance Assessment Task Types
Presenter: Richard Shavelson, Stanford University

A Forum for Addressing Issues of Participative Democracy, Identity, Diversity, and Multiculturalism Through Citizenship Education in Europe: Curricular Implications and International Dimensions
10:35 a.m.-12:05 p.m. Sheraton, Bayside C, 4th floor
Discussant: Heinrich Mintrop, UCLA

The Next Wave: Developing Accountability Systems That Measure What Counts
10:35 a.m.-12:05 p.m. Le Meridien, France II, 3rd floor
Chair: Eva L. Baker, UCLA
Classroom Assignments as Indicators of Instructional Quality
Presenters: Lindsay Clare Matsumura, Helen Garnier, and Jenny Pascal, UCLA
CRESST W~A~V~E~S

CRESST Conference 2002

The CRESST conference will be held Tuesday, September 10, through Wednesday, September 11, 2002 in Los Angeles. Registration materials and complete details about the conference are available on the CRESST Web site, http://www.cse.ucla.edu. There is a significant discount for early registration. (Contact Mary Wilby, wilby@cse.ucla.edu, 310-794-9148, or Kim Hurst, hurst@cse.ucla.edu, 310-794-9140, for registration questions.)

Although topics and sessions have not yet been finalized, we expect many of the following CRESST partners and others to share their work with conference participants.

<table>
<thead>
<tr>
<th>CRESST Co-Director Eva L. Baker, UCLA</th>
<th>Thomas Kane, UCLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRESST Co-Director Robert L. Linn, University of Colorado at Boulder</td>
<td>Haggai Kupermintz, University of Colorado at Boulder</td>
</tr>
<tr>
<td>CRESST Co-Director Joan L. Herman, UCLA</td>
<td>John Lee, UCLA</td>
</tr>
<tr>
<td>CRESST Associate Director Daniel Koretz, Harvard University</td>
<td>Lindsay Clare Matsumura, UCLA</td>
</tr>
<tr>
<td>Jamal Abedi, UCLA</td>
<td>Richard Mayer, University of California, Santa Barbara</td>
</tr>
<tr>
<td>Zenaida Aguirre-Muñoz, UCLA</td>
<td>Heinrich Mintrop, UCLA</td>
</tr>
<tr>
<td>Alison Bailey, UCLA</td>
<td>Robert Mislevy, University of Maryland</td>
</tr>
<tr>
<td>Randy Bennett, Educational Testing Service</td>
<td>Derek Mitchell, UCLA</td>
</tr>
<tr>
<td>Hilda Borko, University of Colorado at Boulder</td>
<td>David Niemi, UCLA</td>
</tr>
<tr>
<td>Christy Kim Boscardin, UCLA</td>
<td>Harry O’Neil, University of Southern California</td>
</tr>
<tr>
<td>Frances Butler, UCLA</td>
<td>Meredith Phillips, UCLA</td>
</tr>
<tr>
<td>Eva Chen, UCLA</td>
<td>Lauren Resnick, University of Pittsburgh</td>
</tr>
<tr>
<td>Greg Chung, UCLA</td>
<td>David Rogosa, Stanford University</td>
</tr>
<tr>
<td>Jonathan Furner, UCLA</td>
<td>Maria Ruiz-Primo, Stanford University</td>
</tr>
<tr>
<td>Robert Glaser, University of Pittsburgh</td>
<td>Michael Seltzer, UCLA</td>
</tr>
<tr>
<td>Pete Goldschmidt, UCLA</td>
<td>Richard Shavelson, Stanford University</td>
</tr>
<tr>
<td>Noelle Griffin, UCLA</td>
<td>Lorrie Shepard, University of Colorado at Boulder</td>
</tr>
<tr>
<td>Edward Haertel, Stanford University</td>
<td>Brian Stecher, RAND</td>
</tr>
<tr>
<td>Brian Junker, Carnegie Mellon University</td>
<td>Yeow Meng Thum, UCLA</td>
</tr>
<tr>
<td>Mary Wilby, Design/Layout</td>
<td>Noreen Webb, UCLA</td>
</tr>
</tbody>
</table>

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002-01, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this publication do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.