Comparability: An Elusive Goal

President George W. Bush’s education plan raises anew some old questions about the comparability of test results from different states. The President’s plan would require states receiving Title I funding to test students every year in Grades 3 through 8. Future Title I funding would be tied to student test score improvement. However, since states use different tests, comparisons of either absolute performance or gains in achievement are problematic. This is recognized, at least implicitly, in the President’s plan and addressed, in part, by the proposal to have NAEP administered annually at the state level in reading and mathematics at Grades 4 and 8. The state trends on NAEP would be used as a check on trends reported by states on their own tests.

Issues of comparability of test results from state to state or district to district are not new. Nor is it new for the federal government to play a role in influencing state testing policies or in trying to achieve comparability across states. Indeed, the federal government has had a fairly substantial impact on testing policies of states and districts at least since 1965 when the Title I program was established to provide funds for compensatory education programs. Title I requirements included the use of tests to provide information that could be used to evaluate the effectiveness of the programs in improving achievement of students served by Title I.

Early Title I reporting was done using grade-equivalent (GE) scores produced from the administration of standardized tests. However, it was quickly decided that the GE scores were not comparable from publisher to publisher, from grade to grade, or from one subject area to another. A gain of one grade-equivalent unit in a year might correspond to a substantial improvement in the percentile rank of a student on one test, maintaining a constant percentile rank on another, or a reduction in percentile rank on a third test. The typical grade-equivalent gains were also quite different for initially high scoring students than for initially low scoring students than for initially low
scoring students. Thus, the commonsense expectation of one GE gain in a year’s time could be unreasonably demanding for some students while too lenient for others. These problems with GEs made the aggregation of results problematic.

In the 1970s, the Title I Evaluation Reporting System (TIERS) instituted a new approach, still based on the results of norm-referenced tests. Reporting gains however, switched to Normal Curve Equivalent (NCE) scores. NCE scores are just transformations of percentile ranks to scores with better analysis properties than percentile ranks.

The idea behind the use of NCE scores was that a student’s percentile rank should remain about the same from one year to the next, given an average year of instruction. Increases in student percentile ranks, transformed to NCE scores, could therefore be reasonably attributed to the positive effect of Title I programs. TIERS therefore attributed positive contributions to Title I if students in the program improved average NCE scores from fall to spring or from one year to the next. Comparability was again in doubt, however, because of questions about the comparability of different standardized tests. The difficulty of the norms from different publishers was highly variable, for example.

These uncertainties about comparability led the U.S. Office of Education to contract for a major study to equate tests of different publishers used for Title I evaluations. The resulting Anchor Test Study equated eight reading tests from six different publishers in Grades 4, 5, and 6. Preliminary review of the content of the most widely used reading and mathematics tests led to the conclusion that the reading tests at those grades were sufficiently similar in terms of the content and skills tested that they could be statistically equated. The study’s authors found, however, that the mathematics tests differed too much in content and skills tested for equating to be meaningful or feasible.

The Anchor Test Study was a remarkable success from a technical standpoint. However, within a few years after the study was completed, all tests included in the study had been replaced by new editions or phased out of use. Thus, comparability that was made possible in one subject at three grade levels was soon undermined. Furthermore, there was strong evidence that even with a single test, there was poor comparability between districts or states that used a fall-to-spring testing cycle to assess NCE gains and ones that used a spring-to-spring testing cycle.

TIERS continued to use NCE score gains despite doubts about the comparability across tests and administration patterns. However, neither the states nor the federal government made much use of the aggregated scores. National Title I evaluations depended on special data collections rather than the aggregation of TIERS results.

In the 1980s when Terrel Bell was Secretary of Education, the U.S. Department of Education started publishing “Wall Charts” of statistics comparing states in terms of average scores on the Scholastic Aptitude Test (SAT) and the American College Test (ACT) along with other statistics. The Wall Charts were harshly criticized because the SAT and ACT were taken only by self-selected samples of students planning to go to college. Further, the proportions of students taking one test or the other varied greatly from state to state depending on tradition and which of the two tests state colleges and universities required for admission. The U.S. Department of Education soon abandoned the Wall Charts. Nonetheless, the Wall Charts accelerated the movement toward finding external benchmarks for comparing student achievement in different states.
Congratulations to all CRESST partners who presented at the recent AERA/NCME meetings in Seattle. Below is a partial list of CRESST partner presentations including both presenters and discussants.

<table>
<thead>
<tr>
<th>Name</th>
<th>Presentation</th>
<th>Session No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abedi, Jamal</td>
<td>Towards Effective Assessment of English Language Learners Through Testing Accommodations and a Portfolio System</td>
<td>AERA 21.19</td>
</tr>
<tr>
<td>Abedi, Jamal</td>
<td>Validity of Standardized Achievement Tests for English Language Learners</td>
<td>NCME A2</td>
</tr>
<tr>
<td>Abedi, Jamal</td>
<td>Validity Considerations in the Assessment of LEP Students Using Standardized Achievement Tests</td>
<td>NCME B2</td>
</tr>
<tr>
<td>Baker, Eva</td>
<td>Implications for Assessment Design</td>
<td>AERA 23.20</td>
</tr>
<tr>
<td>Baker, Eva</td>
<td>Holding Accountability Systems Accountable: Research-Based Standards</td>
<td>AERA 41.23</td>
</tr>
<tr>
<td>Baker, Eva</td>
<td>Technology Principles for the Design and Use of Educational Information</td>
<td>AERA 44.38</td>
</tr>
<tr>
<td>Borko, Hilda</td>
<td>From Classroom to Teaching Policy and Back Again</td>
<td>AERA 9.30</td>
</tr>
<tr>
<td>Borko, Hilda</td>
<td>Collaborative Inquiry for Teacher Leadership: Implications for Teacher Leadership</td>
<td>AERA 13.58</td>
</tr>
<tr>
<td>Borko, Hilda</td>
<td>Schools in Transition: Reform Efforts in Exemplary Schools of Washington</td>
<td>AERA 50.49</td>
</tr>
<tr>
<td>Chen, Eva</td>
<td>Computer Use in Home School Families</td>
<td>AERA 3.23</td>
</tr>
<tr>
<td>Clare, Lindsay</td>
<td>Opening the Black Box in School Reform Evaluation: Investigating the Use of Teachers’ Assignments as Indicators of Classroom Practice</td>
<td>AERA 44.23</td>
</tr>
<tr>
<td>Dietel, Ronald</td>
<td>Comparison of Historical Educational Research Use Issues</td>
<td></td>
</tr>
<tr>
<td>Glaser, Robert</td>
<td>Foundations for a New Science of Assessment</td>
<td>AERA 23.20</td>
</tr>
<tr>
<td>Goldschmidt, Pete</td>
<td>Measuring Determinants of Opportunity to Learn</td>
<td>AERA 41.37</td>
</tr>
<tr>
<td>Huang, Denise</td>
<td>Evaluation and Technology for After School Programs</td>
<td>AERA 5.63</td>
</tr>
<tr>
<td>Herman, Joan</td>
<td>Holding Accountability Systems Accountable: Research-Based Standards</td>
<td>AERA 41.23</td>
</tr>
<tr>
<td>Herman, Joan</td>
<td>Assessment Issues With Special Populations</td>
<td>AERA 50.52</td>
</tr>
<tr>
<td>Koretz, Daniel</td>
<td>Holding Accountability Systems Accountable: Research-Based Standards</td>
<td>AERA 41.23</td>
</tr>
</tbody>
</table>
In the wake of the Wall Charts experience, Congress passed legislation allowing the use of state-level NAEP results for state-by-state comparisons. Since the first administration of state-level NAEP in 1990, a substantial majority, albeit not all, of the states have participated in state NAEP. At two-year intervals, NAEP assesses students in Grades 4 and 8, usually in reading and mathematics.

When George W. Bush’s father was president, the senior President Bush proposed that a set of tests, called the American Achievement Tests, be developed based on the NAEP content. States and districts could use the American Achievement Tests to provide parents and the public with a common performance yardstick. Widespread use of such tests would, of course, also provide a basis for comparing states on that same yardstick. The proposal for the American Achievement Tests was never put into action due to congressional resistance, but the proposal illustrates the long-standing desire to find or develop comparable measures of student achievement across states.

In his proposed Voluntary National Tests (VNT), President Clinton changed the name as well as the particulars for a set of comparable national tests. The VNT proposal reduced the tested content areas to just two—reading at Grade 4 and mathematics at Grade 8—but retained the reliance on NAEP as the framework. Like the American Achievement Tests, the proposed VNT met resistance in Congress based on infringement of local control, potential improper use, and the possibility that national tests might serve as a de facto national curriculum.

As an alternative to the VNT, Congress proposed that different state tests might somehow be equated or linked so that comparisons could be made. The National Academy of Sciences formed a committee to study the feasibility of linking different state tests. After careful analysis, the committee concluded that achieving comparability across the full array of commercial and state achievement tests through linking was not feasible. Both the variations in content and skills assessed and variations in the administration conditions and performance incentives were judged to be too great to make it possible to produce comparable results.

Since achieving the comparability goal was so important, Congress asked the National Academy of Sciences to study the feasibility of embedding part of NAEP into state assessments. Again, the goal was to produce accurate national comparisons. The resulting committee reached the conclusion that the differences in the content, format, and administration of state tests and the material to be embedded from NAEP were so large that valid comparisons could not be achieved though embedding a subset of NAEP items in state and commercial tests.

Comparability of state test results appears to be an elusive goal, but one that is not diminished in its appeal at the federal level regardless of the party in the White House. Resistance by governors and Congress to attempts to impose uniform testing on states so that the goal can be achieved also crosses party lines. It remains to be seen how the latest round of debates about ways of achieving the goal will play out, but the arguments are apt to have a familiar ring.
<table>
<thead>
<tr>
<th>Name</th>
<th>Presentation</th>
<th>Session No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Koretz, Daniel</td>
<td>Toward a Framework for Evaluating Gains on High-Stakes Tests</td>
<td>NCME D1</td>
</tr>
<tr>
<td>Kupermintz, Haggai</td>
<td>Construct Validity Considerations in TVAAS (Tennessee Value Added Assessment System)</td>
<td>NCME D1</td>
</tr>
<tr>
<td>Leon, Seth</td>
<td>Language Fluency Classification Issues</td>
<td>NCME A2</td>
</tr>
<tr>
<td>Linn, Robert</td>
<td>Linking and Equating</td>
<td>AERA 39.25</td>
</tr>
<tr>
<td>Linn, Robert</td>
<td>Holding Accountability Systems Accountable: AERA 41.23</td>
<td></td>
</tr>
<tr>
<td>Linn, Robert</td>
<td>Consequential Validity of Philadelphia’s Accountability System: Triangulating Four Years of Multiple Sources of Evidence</td>
<td>AERA 44.18</td>
</tr>
<tr>
<td>Linn, Robert</td>
<td>Teacher Effects as a Measure of Teacher Effectiveness:</td>
<td>NCME D1</td>
</tr>
<tr>
<td>Mirocha, James</td>
<td>Reliability Issues in Using Standardized Achievement Tests With English Language Learners</td>
<td>NCME A2</td>
</tr>
<tr>
<td>Mislevy, Robert</td>
<td>How to Create Complex Measurement Models: A Case Study of Principled Assessment Design</td>
<td>NCME J1</td>
</tr>
<tr>
<td>Mislevy, Robert</td>
<td>Model Criticism of Bayesian Networks With Latent Variables</td>
<td>NCME Y1</td>
</tr>
<tr>
<td>O’Neil, Harold</td>
<td>Reliability and Validity of a Teamwork Questionnaire</td>
<td>AERA 46.28</td>
</tr>
<tr>
<td>O’Neil, Harold</td>
<td>On-the-Fly Generative Adaptive Assessment of Quantitative Reasoning: Content, Psychometric and Technological Perspectives—Symposium</td>
<td>NCME W1</td>
</tr>
<tr>
<td>Quigley, Denise</td>
<td>Teachers: Training and Evaluation</td>
<td>AERA 11.64</td>
</tr>
<tr>
<td>Shavelson, Richard</td>
<td>Contributions to Generalizability Theory: Indirect But Strong Effects Through Mentoring</td>
<td>AERA 39.25</td>
</tr>
<tr>
<td>Shepard, Lorrie</td>
<td>Three Blueprints for a Revolution: How to Halt the Harm Caused by High-Stakes Tests</td>
<td>AERA 37.27</td>
</tr>
<tr>
<td>Shepard, Lorrie,</td>
<td>Teacher Effects as a Measure of Teacher Effectiveness:</td>
<td>NCME D1</td>
</tr>
<tr>
<td>Stecher, Brian</td>
<td>Schools in Transition: A Statewide Profile of the Impact of Washington’s Educational Reform</td>
<td>AERA 50.49</td>
</tr>
<tr>
<td>Stecher, Brian</td>
<td>Rewarding Schools Based on Gains: It’s All in How You Throw the Dice</td>
<td></td>
</tr>
<tr>
<td>Webb, Noreen</td>
<td>Effects of Group Composition on High-Ability Students’ Achievement in Science</td>
<td>AERA 44.33</td>
</tr>
</tbody>
</table>
Recent CRESST Reports

The following reports are available from the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). To order any CRESST publication or for a complete catalog, contact Kim Hurst at 310-794-9140; email, kim@cse.ucla.edu; or CRESST/UCLA, GSE&IS Building, Mailbox 951522, Los Angeles, CA 90095-1522. Most CRESST publications are also available at no cost on the CRESST Web site, www.cse.ucla.edu.

The Performance of Students With Disabilities on New York’s Revised Regents Comprehensive Examination in English
Daniel Koretz and Laura Hamilton

The Design and Evaluation of Educational Assessment and Accountability Systems
Robert L. Linn
CSE Technical Report 539, 2001 ............$4.50

Making Sense of Data From Complex Assessments
Robert J. Mislevy, Linda S. Steinberg, Russell G. Almond, F. Jay Breyer, and Lynn Johnson
CSE Technical Report 538, 2001 ............$4.50

NAEP Math Performance and Test Accommodations: Interactions With Student Language Background
Jamal Abedi, Carolyn Hofstetter, Eva Baker, and Carol Lord
CSE Technical Report 536, 2001 ............$6.50

Lessons Learned in Using Data to Support School Inquiry and Continuous Improvement: Final Report to the Stuart Foundation
Joan Herman and Barry Gibbons

Leverage Points for Improving Educational Assessment
CSE Technical Report 534, 2001 ............$4.50

An Analysis of Notebook Writing in Elementary Science Classrooms
Gail P. Baxter, Kristin M. Bass, and Robert Glaser
CSE Technical Report 533, 2000 ............$4.00

Using Teachers’ Assignments as an Indicator of Classroom Practice
Lindsay Clare
CSE Technical Report 532, 2000 ............$6.00

On the “Exchangeability” of Hands-On and Computer-Simulated Science Performance Assessments
Anders Rosenquist, Richard J. Shavelson, and Maria Araceli Ruiz-Primo
CSE Technical Report 531, 2000 ............$3.50

Parents and Teachers Working Together to Support Third-Grade Achievement: Parents as Learning Partners (PLP) Findings
Denise D. Quigley & PLP Evaluation Team
CSE Technical Report 530, 2000 ............$6.50

Validating Standards-Referenced Science Assessments
Bokhee Yoon and Michael J. Young
CSE Technical Report 529, 2000 ............$3.50

Standards, Assessments—and What Else? The Essential Elements of Standards-Based School Improvement
Diane J. Briars and Lauren B. Resnick
CSE Technical Report 528, 2000 ............$4.50

Evaluation Report/UC Diagnostic Writing Service
Ellen Osmundson and Joan L. Herman
CSE Technical Report 527, 2000 ............$6.00

Learning to Write in Urban Elementary and Middle Schools: An Investigation of Teachers’ Written Feedback on Student Compositions
Lindsay Clare and Rosa Valdés
CSE Technical Report 526, 2000 ............$3.50
Recent CRESST Reports

The Effects of the Washington State Education Reform on Schools and Classrooms
Brian M. Stecher, Sheila L. Barron, Tammi Chun, and Karen Ross
CSE Technical Report No. 525, 2000 .......... $7.00

The Impact of Preference for Accommodations: The Performance of English Language Learners on Large-Scale Academic Achievement Tests
Martha Castellon-Wellington
CSE Technical Report 524, 2000 ............... $4.00

Instructional Variation and Student Achievement in a Standards-Based Education District
Lauren B. Resnick and Michael Harwell
CSE Technical Report 522, 2000 ............. $3.75

Practical Deliberation in Local School Districts: A South Carolina Experiment
Lorraine M. McDonnell and M. Stephen Weatherford
CSE Technical Report 520, 2000 .............. $5.50

Student Assessment and Student Achievement in the California Public School System
Joan L. Herman, Richard S. Brown, and Eva L. Baker
CSE Technical Report 519, 2000 ............. $6.25

Bayes Nets in Educational Assessment: Where Do the Numbers Come From?
Robert J. Mislevy, Russell G. Almond, Duanli Yan, and Linda S. Steinberg
CSE Technical Report 518, 2000 ............. $4.00

The Role of Classroom Assessment in Teaching and Learning
Lorrie A. Shepard
CSE Technical Report 517, 2000 ............. $7.00

Assessing Active Knowledge
Robert Glaser and Gail P. Baxter

The Dependability and Interchangeability of Assessment Methods in Science
Noreen M. Webb, Jonah Schlackman, and Brenda Sugrue
CSE Technical Report 515, 2000 ............. $4.00

“No Excuses”: School Reform Efforts in Exemplary Schools of Kentucky
Shelby Wolf, Hilda Borko, Monette McIver, and Rebekah Elliott

Developing Indicators of Classroom Practice to Monitor and Support School Reform
Pamela R. Aschbacher

Professional Development: A Key to Kentucky’s Reform Effort
Hilda Borko, Rebekah Elliott, and Kay Uchiyama

Identifying Differential Item Functioning on the NELS:88 History Achievement Test
Vi-Nhuan Le
CSE Technical Report 511, 1999 .......... $2.50

Accuracy of Year-1, Year-2 Comparisons Using Individual Percentile Rank Scores: Classical Test Theory Calculations
David Rogosa

Accuracy of Individual Scores Expressed in Percentile Ranks: Classical Test Theory Calculations
David Rogosa
CSE Technical Report 509, 1999 .......... $5.50

Accommodations for Students With Disabilities: A Teacher’s Guide
Ann M. Mastergeorge and Judy N. Miyoshi
CSE Technical Report 508, 1999 .......... $4.00

Knowledge Mapping in the Classroom: A Tool for Examining the Development of Students’ Conceptual Understandings
Ellen Osmundson, Gregory K. W. K. Chung, Howard E. Herl, and Davina C. D. Klein

State Standards-Setting and Public Deliberation: The Case of California
Lorraine M. McDonnell and M. Stephen Weatherford
CSE Technical Report 506, 1999 .......... $5.00

Spring 2001
Baker Receives UCLA Alumni Award

The UCLA Alumni Association recently honored CRESST Co-director Eva L. Baker with the Professional Achievement Award, established in 1962 to recognize superior achievements by UCLA alumni. Baker was honored for her many contributions to the educational community, including scholarship and teaching. Among her accomplishments are the design and development of a new standards-based assessment system, a data analysis program called the Quality School Portfolio to help schools analyze and use data, and her service as co-chair of the committee that revised the Standards for Educational and Psychological Testing in 1999.

A recipient of the American Educational Research Association’s Presidential Citation and the California Educational Research Association’s Lifetime Achievement Award, Baker chairs the National Research Council’s Board on Testing and Assessment.

Linn Selected as AERA President-Elect

Herman is AERA Program Chair-Elect

CRESST Co-director Robert L. Linn was recently selected as the president-elect of the American Educational Research Association. He will begin his presidency of the 23,000 member research association at the conclusion of the 2002 annual AERA meeting in New Orleans. A widely recognized measurement and testing scholar, Linn will be supported in his duties by CRESST co-director Joan Herman who will serve as the AERA program chair. In her role, Herman will be responsible for the overall management of approximately 1300 sessions and 4000 presentations. Our congratulations to both Bob and Joan!

Anne Anastasi Mourned

The measurement community mourns the loss of Anne Anastasi, a pioneer woman in the testing and psychology field. She passed away on May 4, 2001, in New York City at the age of 92. Anastasi was the author of many well-known publications including Psychological Testing, one of the most widely respected books in the field. A former president of the American Psychological Association, she received the National Medal of Science in 1987.