POLICY BRIEF 5

CRESST

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

Winter 2002

Clockwise from top: Daniel Koretz, Eva L. Baker, Robert L. Linn, and Joan L. Herman.

Standards for Educational Accountability Systems

CRESST Co-Directors Eva L. Baker, Robert L. Linn, Joan L. Herman, and CRESST Associate Director Daniel Koretz

The Standards for Educational Accountability Systems is a collaborative project between CRESST and the Consortium for Policy Research in Education (CPRE). The standards reflect input from Eva L. Baker, Robert L. Linn, Joan L. Herman, Daniel Koretz, and Richard Elmore, as well as reviewers from professional organizations, educational institutions, and commercial test producers. The Standards will appear in a forthcoming book edited by Susan Fuhrman and Richard Elmore, *Redesigning Accountability Systems* (New York: Teachers College Press).

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?

Now is the time for states, as they reflect and prepare for action, to consider anew the true quality of their future efforts.

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* (AERA, APA, & NCME, 1999), research findings on testing and accountability systems, and studies of best practices. The Standards for Educational Accountability Systems also stress the importance of understandable description of accountability systems and clear 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 its 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 consider-

ations. 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. Standards on System Components

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

   Comment: Explicit information about expectations is a prerequisite for participants to perceive the accountability system as fair. It is also needed to allow participants to meet expectations and to monitor their progress.

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

   Comment: Although measures of student achievement may be of primary interest for accountability purposes, it is important also to obtain information about student and teacher characteristics to provide context for interpreting student achievement. It also is important to consider other student outcome data such as attendance, mobility, and rates of retention in grade, dropout and graduation. Moreover, it is important to obtain data on instructional resources and curriculum materials, and about the degree to which students are provided with adequate opportunity to learn the content specified in content standards and curriculum materials.

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

   Comment: Students, teachers, administrators, and policymakers have a shared responsibility for achieving the results expected by accountability systems. The system needs to provide the information for each of these parties to know what actions need to be taken.

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

   Comment: Previous practices that excluded many students from testing because of absence on the day of test administration, limited English proficiency, or student disabilities gave a distorted and usually exaggerated view of overall performance. They also precluded accountability for the performance of excluded students. Legal requirements as well as ethical considerations demand that all students be included in the accountability system. Many students who would have been excluded in the past can be included without any alterations in the test or administration conditions. Some accommodations in administration conditions will be required for other students, and for some students the test will need to be modified, or alternative assessments used, in order for the students to be included in the accountability system. No student should be left out of the system, however.

> Many students who would have been excluded in the past can be included without any alterations in the test or administration conditions.

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

   Comment: Making sense of overall accountability indices requires an understanding not only of the elements that go into the index, but of the weights that are assigned to each element. It is informative to provide not only the weights that are assigned to the different elements by policy, but also information about how much each element affects the overall index. The relationship of an element to a weighted accountability index depends on the variability of the element across institutions as well as the weight assigned to the element by policy.

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.

Comment: Progress based on student test averages reflecting performance of the total or subgroup is usually not regular because of changes in school populations, measurement error, and other situational factors. Approaches that capture the longitudinal performance of individuals (along with an indicator of the proportion such longitudinal data represent) can help minimize inappropriate inferences. Other strategies include using more than one year's difference to compute growth.

B. Testing Standards

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

Comment: There are several reasons for this standard. First, no test is perfectly reliable. There is always a degree of uncertainty associated with any test score. That uncertainty needs to be taken into account when making decisions about individual students. Second, all tests have less than perfect validity. Hence, it is important to look for other information that will either support or disconfirm the information provided by a single test score. The importance of obtaining other information to confirm or disconfirm the information provided by a single test score increases as the importance of the decision and the stakes associated with it increases. Yet another reason for multiple sources of information is the limitation of a single measure as a sample of the domain(s) of interest.

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

Comment: The items contained on a test form are only a sample of the domain that the test is intended to measure. Learning the answers to the items on a single form by focusing exclusively on those items is not the same as learning the material for the domain of content the test is intended to measure. Consequently, it is important to evaluate the generalizability of performance by administering a different form when a test is administered for a second or third time.

The importance of obtaining other information to confirm or disconfirm the information provided by a single test score increases as the importance of the decision and the stakes associated with it increases.

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.

Comment: Validity is dependent on the specific uses and interpretations of test scores. It is inappropriate to assume that a test that is valid when used for one purpose will also be valid for other uses or interpretations. Hence, validity needs to be specifically evaluated and documented for each purpose.

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.

Comment: Tests need to be sensitive to differences in instructional quality and student effort in order to be useful as tools in improving system performance. Sensitivity to instruction and to student effort is also a prerequisite for fairness if educators and students are to be held accountable for results.

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.

Comment: Because tests are fallible measures, classification errors are inevitable when tests are used to classify students or institutions into categories associated with rewards or sanctions. In order to judge whether the risk of errors is acceptably low, it is essential that information be provided about the probability of misclassifications of various kinds.

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

Comment: Validity needs to be assessed separately for students with different language backgrounds. Whether a test is administered in English or in a student's primary language, validity of the test for students of different language backgrounds cannot be assumed from evidence based only on test results of students whose first language is English. Testing students in their primary language may be required for some students. However, translation and adaptation of tests to different languages

is a complex undertaking. There are many threats to validity of tests administered in different languages. Lack of consistency between the language of the test and the language of instruction is one of the major threats to validity and needs to be evaluated to the extent feasible.

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

Comment: Accommodations may be needed for some students with disabilities to be able to participate in testing in a meaningful way. The goal of accommodations is to remove sources of difficulty that are irrelevant to the intent of the measurement. That is, an accommodation should make it possible for a student with disabilities to demonstrate her knowledge and skills in the content domain being tested so that the score reflects that knowledge and skill rather than the student's disability. The accommodation should level the playing field; it is not intended to give the student with a disability an advantage over other students. The validation task is to provide evidence that the test is reflecting the student's knowledge and skills and not her specific disability. For students with severe disabilities, assessments may need to be modified, or alternative assessments may need to be selected or developed, possibly designed to assess different learning goals than those of the assessments used for the majority of students. Evidence regarding the validity of interpretations made from modified or alternative assessments should be provided to the extent feasible.

The accommodation should level the playing field; it is not intended to give the student with a disability an advantage over other students.

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.

Comment: The degree of alignment of a test with content standards may be evaluated, for example, by providing a mapping of the test specifications to the content standards. Such a mapping can reveal areas of the content standards that are not included in the test specifications as well as areas that are lightly or heavily sampled in the test specifications. The mapping may also reveal areas tested that are not part of the content standards. Performance standards generally provide verbal descriptions of performance levels that are considered satisfactory or exemplary. The degree to which the descriptions map directly to the test items and the correspondence of the performance standards to the cut scores on the test need to be documented and evaluated.

C. Stakes

15. Stakes for accountability systems should apply to adults and students and should be coordinated to support system goals.

Comment: Asymmetry in stakes may have undesirable consequences, both perceived and real. For example, if teachers and administrators are held accountable for student achievement but students are not, then there are likely to be concerns about the degree to which students put forth their best effort in taking the tests. Conversely, it may be unfair to hold students accountable for performance on a test without having some assurance that teachers and other adults are being held accountable for providing students with adequate opportunity to learn the material that is tested. Incentives and sanctions that push in opposite directions for adults and for students can be counterproductive. They need to be consistent with each other and with the goals of the system.

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

Comment: Extenuating circumstances may call the validity of results into question. For example, a disturbance during test administration may invalidate the test results. Also, individuals may have information that leads to conflicting conclusions about performance. Appeal procedures allow for such additional information to be brought to bear on a decision and thereby enhance its validity.

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

Comment: Making plans for phasing in stakes for results is part of making accountability expectations explicit to participants. Explication of plans allows participants to make informed decisions about how best to achieve the ends expected by the accountability system.

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

Comment: Starting with broad, diffuse stakes (e.g., public reporting of aggregate achievement results for schools) allows participants time to make the changes needed to meet expectations before being confronted with specific rewards or sanctions for performance (e.g., monetary rewards to schools or teachers, graduation requirements for students). Advance warning and phasing-in of stakes enhances both the perception of fairness and the actual fairness of the accountability system.

Advance warning and phasing-in of stakes enhances both the perception of fairness and the actual fairness of the accountability system.

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.

Comment: The press plays an important role in the interpretation of the results produced by accountability systems. Legitimate interpretations of results require an understanding of what goes into them and some of their technical characteristics. Those responsible for the accountability system also have a responsibility to help ensure proper interpretation of the results and to minimize inappropriate interpretations to the extent possible. Efforts to assist the press in understanding the results, their strengths and limitations, and the legitimate and illegitimate interpretations can pay considerable dividends in improved coverage by the press and better understanding by the public.

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 subgroups.

Comment: Interpretations of results can be enriched by the reporting of consistencies and inconsistencies provided by multiple indicators of performance. Performance by subgroups needs to be considered to ensure that overall results do not conceal great disparities in subgroup performance. Understanding the degree of uncertainty in results can reduce the likelihood of misinterpretation and enhance the likelihood of appropriate use of results.

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, recruiment, and retention; and
g. produces unanticipated outcomes.

Comment: The primary purpose of educational accountability systems is to improve instruction and student learning. The overarching evaluation question is the degree to which the intended benefits are realized and the costs in terms of unintended negative consequences are minimized. Listed items (a) through (d) reflect intended positive consequences the realization of which is the focus of evaluation. Items (e) and (g) emphasize the needed evaluation of plausible unintended negative consequences. Item (f) requires the evaluation of both intended positive and unintended negative influences of the accountability system.

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.

Comment: Gains in performance may be spurious or real. Evaluation of the gains may be aided by investigations of the degree to which gains on the measures used by the accountability system are reflected in changes on alternative indicators of performance obtained from other tests or more general indicators, such as performance beyond school in college or the workplace. Differential effects on identifiable student or personnel groups may lead to different conclusions than those that are supported by the overall aggregate performance.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Note: *An abbreviated version of this Policy Brief appeared in CRESST Line, Winter 2002.*

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. To order copies of this policy brief, contact Kim Hurst, 310-794-9140, email: kim@cse.ucla.edu, or write to Kim at CRESST/UCLA, GSE&IS Building, Mailbox 951522, Los Angeles, CA 90095-1522.

GSE&IS Logo

UCLA Center for the Study of Evaluation
CSE/CRESST
GSE&IS BLDG MAILBOX 951522
LOS ANGELES CA 90095-1522
ADDRESS SERVICE REQUESTED
www.cse.ucla.edu

NON PROFIT ORG.
U.S. POSTAGE
PAID
U.C.L.A.

EEX1