

# The CRESST *Line*

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

FROM THE DIRECTORS:

## Minimum Group Size for Measuring Adequate Yearly Progress

CRESST Co-Directors

Robert L. Linn, Eva L. Baker, and Joan L. Herman

The No Child Left Behind Act of 2001 ([NCLB] 2002; Public Law 107-110) requires schools, school districts and states to report the percentage of students scoring at the “proficient level or higher” on state assessments not only for all students as a whole but also for specific groups of students. These groups include “economically disadvantaged students, students from major racial and ethnic groups, students with disabilities, and students with limited English proficiency” (NCLB, 2002).



Joan L. Herman, Eva L. Baker, and Robert L. Linn

However, disaggregated reporting (by group) is *not* required “in the case in which the number of students in a category is insufficient to yield *statistically reliable information* or the results would reveal personally identifiable information about an individual student” (NCLB, 2002, emphasis added). Many states and school districts are struggling in particular with this “minimum group size” issue, which we address in the remainder of this article.

### Background

There are good reasons to report achievement by categories of students. We know that a disproportionate number of students from the categories targeted by NCLB have lagged behind on achievement tests for many years. Disaggregated reporting for those categories of students provides a mechanism for monitoring the degree to which the goal of leaving no child behind is reached.

Disaggregated reporting also helps to monitor progress in closing the achievement gap, another key NCLB goal. Texas’ success in closing the achievement gaps between African American and White students and between Hispanic and White students on the Texas Assessment of Academic Skills (TAAS) has been cited frequently as an example of the usefulness of disaggregated test score reports.

We support the benefits of disaggregated reporting where the reporting can be done in a way that yields “statistically reliable information” and in a way that does not “reveal personally identifiable information about an individual student.” For statewide reports and reports for large districts, the number of students within even the smallest category will usually be sufficient to meet both requirements.

## Inside

From the Directors.....1

CRESST 2002 Conference.....2

Quality School Portfolio—  
Web Version.....3

CRESST *W~A~V~E~S*.....8

## Save the Date

2003 Annual CRESST Conference  
September 4, 5, and 6  
Location: UCLA

[www.cse.ucla.edu](http://www.cse.ucla.edu)

# CRESST 2002 Annual Conference: “Civil Discourse” on Assessment

Part 1, Anne Lewis

As the No Child Left Behind Act of 2001 (NCLB) began its momentous impact on American education, CRESST marshaled its ongoing research agenda to focus the 2002 annual conference precisely on the issues educators and policymakers will confront in the new legislation. For two days, September 10-11, conference participants engaged in what CRESST Co-director Robert Linn called “civil discourse” on many aspects of assessment and accountability.

A statement by the National Research Council, quoted by CRESST Co-director Eva Baker, set the tone for the conference, *Research Goes to School: Assessment, Accountability, and Improvement*. It warned that high stakes should not be applied to any assessment until its validity, reliability, and fairness had been addressed.

Baker opened the first panel discussion by tackling a major validity issue—using tests and assessments to improve instruction. They must be integral to the process of learning, she said, “not just tacked on at the end.” Emphasizing that “knowledge is power,” and that those using tests and assessments should know their qualities well enough to use them powerfully for instruction, she explained why, presently, the barriers to using tests wisely to improve instruction are greater than the incentives. Most tests used for accountability “lightly sample content” and fulfill multiple purposes when “there is not a lot of validity” to support multiple purposes. Ideally, assessment designs should reflect learning beyond what is domain specific, use well-sampled content, contain a process and criteria for open-ended performance, and show evidence on several points such as sensitivity to test preparation versus teaching significant content and intellectual skills.

Another panelist, Dan Koretz of CRESST and Harvard University, elaborated on the last point—the validation of gains on tests that may make the sampled content unrepresentative, even after initial validation. Traditional valida-

tion, he said, is insufficient in a high-stakes environment because it ignores behavioral responses to testing (e.g., coaching) and inadvertent emphases in tests and assumes some stability in the process (e.g., consistency on what is left out). Analyses of Kentucky and Texas test results over time illuminate these issues. Gains on state tests in these states, for

example, are not repeated in scores on the National Assessment of Educational Progress (NAEP). CRESST is developing a new framework for valida-

tion of tests, Koretz said, that includes teacher surveys and interviews, statistical models, identification of substantive and nonsubstantive performance elements in a test, studying how teachers reallocate attention to subjects, and finding out “how teachers use shortcuts.” The study also is examining the nature of test coaching.

The issue of validity begins when the tests are constructed, according to Stephen Dunbar of the University of Iowa and a principal co-author of the Iowa Tests of Basic Skills. The current environment may well be one of “leave no test item behind,” he said. From the test developer’s perspective, the selection of material at the very beginning is critical, though it is seen as the “grunt work” of assessment. Ideally, the process of selection is followed by field testing, then review, revision, and replacement, and a repeat of the process once more. The search for material is arduous, and “if I don’t throw away about half of the initial items developed, I’m not doing my job,” he said. Unfortunately, Dunbar noted, it is at this point in the process where money can be saved. The unprecedented scope of testing under the No Child Left Behind Act will place heavy demands on test development and result in less field testing and perhaps a dwindling quality of the items.

The discussion that followed the panel presentations reflected several concerns about the current assessment environment: Are test

“[I]f I don’t throw away about half of the initial items developed, I’m not doing my job,” Dunbar said.

# Quality School Portfolio—Web Version

Margaret Heritage

The No Child Left Behind Act of 2001 (NCLB) mandates annual testing of students and disaggregated reporting of test data to improve student achievement. Longitudinal analysis of student and subgroup performance can support educational improvement by identifying instructional strengths and weaknesses. But assessment data itself will not necessarily improve instruction and student learning. Educators must be able to analyze data, draw accurate conclusions, and take actions that will promote student learning, either individually or as a group.

To meet these needs, CRESST developed the Quality School Portfolio (QSP), a decision support tool. Funded by the U.S. Department of Education and other sources, the desktop Quality School Portfolio is now used in more than 1,000 schools, 80 districts, and in all 50 states. QSP can be customized to allow schools to add local variables on achievement and other data to meet unique site requirements.

CRESST designers are now creating a Quality School Portfolio—Web Version, with expanded capabilities and increased flexibility. A beta version is complete, and a pilot implementation study begins in January 2003.

Web QSP has many features to help educators understand and use data to improve student achievement, including individual, longitudinal records for each student. Test

Name	Year	Score	Date	Teacher
Lincoln Video	9	A	10/22/01	Michaelson
Booker T Essay	9	proficient	05/22/02	Michaelson
Concept Map 1	10	65	08/04/02	Davis
US Song	10	88	01/12/03	Davis

scores and other data can be disaggregated by various groups and, through the report function, can be transformed into easy-to-understand, action-based graphs for decision making at the district, school, and teacher level.

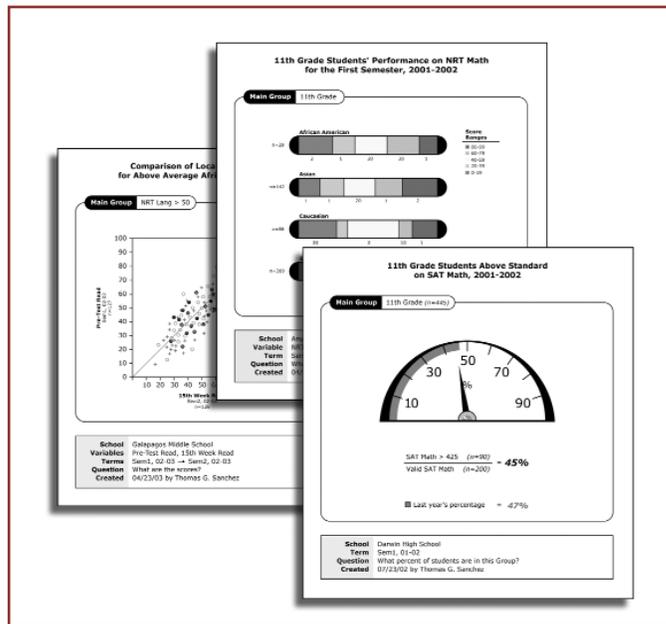
Samples of student work and performance can be stored in the digital portfolio. Using the goal

function, educators can determine goals and set targets to monitor student progress towards meeting standards.

Studies by CRESST researchers have shown that educators using the QSP desktop version increase their use of school data, use data in new ways, and develop a better understanding of student achievement. Web

QSP will make data analysis available to all schools and teachers nationwide, assist schools and districts in meeting NCLB requirements, and increase effective data analyses that support instruction.

For more information, please visit the QSP Web site at [qsp.cse.ucla.edu](http://qsp.cse.ucla.edu).



For small districts and individual schools, however, the number of students in some or many categories can be quite small. Hence, there is a need to consider the minimum number of students in a category that will produce results with sufficient statistical reliability to justify reporting.

Some observers have suggested that because all or nearly all students in a school are tested, sampling error is not a major concern. As Cronbach, Linn, Brennan, and Haertel (1997) have noted, that interpretation is reasonable so long as the school result is viewed simply as a historical description of what happened. However, the interpretation is not reasonable when the result is used, as it always is, as an indicator of school effectiveness. Using student test results to conclude “that a school is effective as an institution requires the assumption, implicit or explicit, that the positive outcome would appear with a student body other than the present one, drawn from the same population” (Cronbach et al., 1997, p. 393). Thus, statistical information about the variability in observed results due to sampling is relevant in setting the minimum number of students in any group.

One thing is clear at the outset. There is no magic number below which results would have zero statistical reliability and above which they would have excellent statistical reliability. Rather, statistical reliability increases gradually and steadily as the number of students increases—or, more precisely, as a function of the square root of the number of students. The statistical reliability of a percentage based on 100 students will be twice as good as that of a percentage based on 25 students. The challenge is to set a standard that will yield results with sufficient statistical reliability to appropriately hold schools accountable. Thus, the goal in establishing a minimum number of students is analogous to what is sometimes referred to as the “Goldilocks” standard. The minimum number should not be set so high that the potential benefits of disaggregated reporting are lost, but neither should it be set so low that there is an unacceptably high probability that schools will receive sanctions as the result of random fluctuations for students in a low-frequency category.

The minimum size issue is even more

important because NCLB will require school sanctions if even a single group fails to show sufficient progress. Thus, a school that meets its “adequate yearly progress” (AYP) target for students as a whole

and for all groups except one could be sanctioned. Schools missing their target in a single category are likely to redouble their instructional effort with those students. Such an outcome could produce the intended impact of improving the achievement of students in that category, but other groups might receive less emphasis. Any of the other groups might thereby fail to make adequate yearly progress. Perhaps worse, students might transfer to other, more successful schools, thereby exacerbating the original school’s perceived failure. The possible consequences are so great that measurement accuracy must be a very high priority.

#### Deriving a Minimum Sample Size

We consider a specific situation and specific statistics to address the issue of a desired sample size. The NCLB goal is that 100% of students will be proficient or above on their state tests by the year 2014. States are supposed to set intermediate AYP targets on a straight-line path toward the 100% goal. For example, a state with a baseline 40% proficient or above in 2002 would have to set AYP targets that increased the percent proficient or above by  $(100\% - 40\%)/12 = 5\%$  per year.

At the school level, the percentage of stu-

**Table 1**

The Standard Error of the Difference of Independent Samples as a Function of the Number of Students in Each Sample When the Average of

Number of Students in Each Sample
10
15
20
25
30
35
40
45
50
60
70
80
90
100

dents in any particular category who score at the proficient level or higher is subject to several sources of uncertainty. The first uncertainty is the measurement error on any test, but this

Standard Error of Difference Between Percentages for Two Samples of the Number of Students in the Two Percentages is 50

Standard Error of Difference in Percentages
22.4
18.3
15.8
14.1
12.9
12.0
11.2
10.5
10.0
9.1
8.5
7.9
7.5
7.1

error contributes much less to the overall uncertainty than the student sample (Cronbach et al., 1997), which is the second and more serious uncertainty.

To determine a minimum sample size, we investigated the standard error of the difference between percentages from two independent samples. Table 1 shows the standard error as a function of student sample size. Even with 50 students in a category each year, the standard error of the difference between the percentage in year 2 and that in year 1 is 10%. The observed difference is expected to be within one standard error of the true difference two

thirds of the time. The observed difference falls outside the boundary of the true difference by more than a standard error one third of the time. Thus, about 1 time in 6 the percentage of students in the sample who are proficient would be no larger in year 2 than in year 1, even when the instruction had improved enough to increase the percentage proficient for an indefinitely large number of students by 10%.

If the number of students in the category was 25 rather than 50, the standard error would be about 14% rather than 10%. It would take an improvement in instruction great enough to produce an increase in the percentage of students in the long run who would score at the proficient level or higher of 14% to reduce the

chances as low as 1 in 6 that the percentage of the 25 students in year 2 who scored at the proficient level would be less than or equal to the corresponding percentage for their 25 counterparts in year 1.

If the minimum number of students in a category was set at 50, the number of groups that would qualify for disaggregated reporting would be relatively small at most schools. Obviously, more groups would qualify for disaggregated reporting if the minimum number were set at 25. With either 25 or 50 students, the identity of individual students would be protected.

Other examples could be considered by changing the average percentage to values other than the 50% used in Table 1. Standard errors will vary for cases where the number of students changes from one year to the next, but the general trend in Table 1 provides a reasonable background for considering the trade-offs between selecting a larger and smaller minimum sample size for disaggregated reporting. More disaggregated reporting will be achieved with a smaller minimum number of students, but a larger minimum number will provide greater protection against mistakenly identifying schools for improvement as the result of the low statistical reliability of the difference in the percentage of students who score at the proficient level or higher from one year to the next.

In our judgment, a reasonable compromise between the competing goals of more disaggregated reporting and greater statistical reliability would be to set the minimum number of students at 25. We recognize, however, that others may make different judgments.

## References

- Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, 57, 373-399.
- No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

results valid reflections of standards, and are standards sometimes less challenging than what teachers require? What happened to the idea of using multiple measures of performance? Aren't the undesirable responses to external testing "cold-blooded reality" for principals? Are the inferences from test scores on school improvement all that important?

The basic issue, Baker said, is that the process to develop and use valid assessments, that is, our "attempt to do good things," is being constrained by time and cost factors. Until a very different construct is developed, "people may have to settle for a very imperfect system."

### Adequate Yearly Progress

The second CRESST panel addressed the central issue of accountability in No Child Left Behind—the definition of "adequate yearly progress" (AYP). Key criteria for states in developing their definitions are stringent and meant to bring all of a state's students up to its standards by 2014. Continuous progress with all subgroups of students must be shown. If a school's starting point (defined by 2001-02 scores) is 52% proficient in reading and 40% in math, then it must show adequate yearly progress of 4% gains in reading and 5% gains in math in order to reach the target, Robert Linn explained in his presentation. One major problem will be the minimum number of students counted in each category, he said. Too low a figure, say 10 students, will give unreliable results because of the standard statistical error. If the minimum were set at 50, the number of subgroups whose scores would need to be disaggregated would be relatively small at most schools.

Using a definition of AYP based on longitudinal data would have certain advantages, Linn said. Individual student growth would become the basis of measurement, and schools would be accountable for only those students who attended the school for a full year. However, the assumption that a cohort score would hold for the following year would not be true in schools with high mobility, he said.

Another panelist on the same subject, Edward Haertel of CRESST and Stanford University, raised questions about how much

adequate yearly progress can be expected. The "safe harbor" provision (if a single subgroup falls below the mark, a school needs to show only 10% improvement for it) makes it easier for high-achieving schools to avoid sanctions, he noted, but even so, few will make the grade.

Are the annual improvement targets reasonable? Haertel noted that if progress were based on the experience of NAEP over the years, it would take 110 years to reach 100% proficiency across the country. A simulation of AYP under California's standards shows that virtually all schools would be classified as needing improvement by the time of the NCLB deadline. A second run on the simulation that looked at the characteristics of subgroups, using racial effects as a proxy for a number of factors, found that a school's progress was very dependent on the economic and educational level of the mother. "Terrible mischief is done when this is not emphasized," he said.

The use of racial subgroup [reporting] rules—as in the California accountability system and in the federal No Child Left Behind Act—can generate unintended consequences, according to Thomas Kane of CRESST and the UCLA School of Public Policy and Social Research. Under the California system, a school is expected to reach growth targets for students overall and for each of its subgroups to win an award. However, subgroup test scores can be quite volatile from year to year, particularly since subgroups can contain as few as 30 students. Suppose that a school was doing equally well with all its subgroups and had a 70% chance of reaching the target for any group. A school with one subgroup would have a 70% chance of winning. But since the annual fluctuations are nearly independent for different subgroups, a school with two subgroups would have a 49% chance, and a school with three subgroups would have a 35% chance.

Therefore, for purely statistical reasons, Kane said, the more subgroups a school has, the less likely it will be able to reach its target. "It is ironic," he said, "that the rules which are intended to help African American and Latino students end up directing fewer resources to such students since they are more likely to attend racially heterogeneous schools than White students."

Even if there are some unintended conse-

quences, a test-based accountability system for schools can be worthwhile, because even small improvements in students' achievement have huge payoffs in terms of subsequent lifetime earnings. However, Kane said, "we need to be careful in designing accountability systems to minimize the perverse incentives, such as those often created by racial subgroup rules."

### **Validity of Accountability Models**

A final panel for the day—discussing the validity of accountability models—featured the latest data from the popular Tennessee value-added model. Statistician William Sanders, who developed the model, asserted that too much was being made of different kinds of tests and not enough of analyzing the data from the tests. The value-added system, now with data from at least one district in each of 21 states using a variety of tests, reveals many aspects of teaching. Teacher effectiveness, for example, definitely relates to teacher experience but not in a continuous line. It improves for 12 years, levels off, then declines after 20 years, he said. A study of Algebra 1 teachers found that those teaching on a provisional basis had 30% effectiveness, whereas those who had gone through the complete licensure process had 40% effectiveness. Middle school math teachers with secondary certification were definitely more effective than those with elementary certification, he said. Sanders said he did not claim that his model is the best, "but the challenge to the research community is to take the complexities such models reveal and extract the most powerful measures of barriers and successes in student achievement."

The theory of action regarding NCLB, according to Brian Stecher of CRESST and RAND, uses standards "as the driving factor." Academic standards lead to assessments and to educational policies and practices that affect student learning. These produce adequate yearly progress or consequences. But critical questions need to be asked about some components of the theory, Stecher said. Are the standards complete, clear, challenging, and useful for instructional planning? Are the assessments aligned, instructionally sensitive, and valid? Are proficiency levels and AYP sensitive to growth, reasonable to achieve, and equitable? Are the consequences

fair, equitable, and persuasive but not coercive?

Stecher also wondered what would affect how the education system changes in response to NCLB, such as how well the provisions are described and transmitted; how teachers, parents, and policymakers respond to scores; and what changes occur in policies, practices, and outcomes. The effects on policymakers could go either way, he added, either allowing them to judge the effectiveness of their policies and fostering better allocation of resources, or encouraging them to adopt policies that raise test scores narrowly, waste resources on test preparation, and distract students from learning.

Describing different models of accountability, Brian Gong of the National Center for the Improvement of Educational Assessment, Inc., called NCLB a model that looks at performance as a point in time. It could provide for other models, but many of the decisions made in NCLB are "bad designs," he said. For example, the probability of a school being misclassified is very high. Moreover, "what characterizes a bad school is not the opposite of what it means to be a good school," he said. "A bad school doesn't teach reading well, but a good school does more than teach reading and math well."

### **Redemption for Psychometric Sins**

UCLA Professor Emeritus W. James Popham received the 2002 CSE/CRESST Distinguished Achievement Award but did not shrink from chastising his colleagues and profession for some "heavy-duty sins" regarding tests in a high-stakes environment. Traditionally constructed tests, or norm-referenced ones, play an important role because the relative information about a student can be useful to teachers and parents. This comparative measurement, however, "is dead wrong for evaluating the quality of instruction," he said. Current use of such tests has led to curricular reductionism, the "jettison of joy" from classrooms because teachers are drilling students for tests, and "modeled dishonesty" on the part of teachers who will do anything to improve test scores.

The educational measurement field, Popham continued, has been guilty of "sins of omission" by not opposing the misuse of tests. Off-the-shelf, nationally standardized tests come from reputable testing firms, so policy-

# CRESST W~A~V~E~S

continued from page 7, CRESST 2002 Annual Conference



On September 10, 2002, CRESST was honored to present this year's CSE/CRESST Distinguished Achievement Award to Professor Emeritus W. James Popham. Professor Popham has made innumerable contributions to the educa-

tional measurement field through his leadership and scholarship. He taught evaluation and measurement courses for most of his career at UCLA's Graduate School of Education while authoring 20 books, 180 journal articles, 50 research reports, and 150 papers. Always known for his sense of humor, in his biography, he claims authorship of 1,426 grocery lists.

In 1978 Professor Popham served as the president of the American Educational Research Association and was the founding editor of *Educational Evaluation and Policy Analysis*. He received the 2002 National Council on Measurement in Education Award for Career Contributions to Educational Measurement.

makers have assumed they were suitable for their accountability purposes, "and by our silence, we let them think it."

For penance, Popham proposed that the psychometric field actively promote increased assessment literacy among policymakers, practitioners, the public, and especially parents of school-age children. A more assessment-literate citizenry, he said, "will be able to push for the kinds of educational tests that not only supply accurate accountability evidence by which educators can be evaluated, but also can support improved instructional practices by those educators." The penances might include writing magazine articles and op-ed essays, making oral presentations, identifying resources for colleagues, establishing state and local assessment literacy councils, enlisting allies such as PTAs, developing materials for assessment literacy, and organizing conferences for citizens to promote assessment literacy. Unless the testing community conducts some form of penance, its members might be condemned to perdition and a lifetime of explaining to bystanders "the nature of construct-related validity." Consider that sort of eternity, he concluded, "and you'll surely choose some sort of penance-paying option."

---

## GRADUATE SCHOOL OF EDUCATION & INFORMATION STUDIES

---

The Part 2 summary of the CRESST 2002 annual conference will appear in the Winter 2003 *CRESST Line* issue.

Center for Research on Evaluation,  
Standards, and Student Testing  
Eva L. Baker, Co-director  
Robert L. Linn, Co-director  
Joan L. Herman, Co-director  
Daniel Koretz, Associate Director  
Ronald Dietel, *CRESST Line* Editor  
Katharine Fry, Editorial Assistant

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

PRESORTED  
STANDARD  
U.S. POSTAGE  
PAID  
PASADENA, CA  
PERMIT NO. 1132

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, 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.