Commonly Unrecognized Error Variance in Statewide Assessment Programs

---

Sources of Error Variance and
What Can Be Done to Reduce Them

Prepared for the
Technical Issues in Large Scale Assessment (TILSA)
State Collaborative on Assessment and Student Standards (SCASS)
of the Council of Chief State School Officers

CCSSO
Council of Chief State School Officers

THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

Commonly Unrecognized Error Variance in Statewide Assessment Programs

A paper commissioned by the
Technical Issues in Large-Scale Assessment State Collaborative
Council of Chief State School Officers

Content Prepared By:
Frank Brockmann, Center Point Assessment Solutions

Supported by TILSA Advisers:
Duncan MacQuarrie
Doug Rindone
Charlene Tucker

Based on Research By:
Gary W. Phillips, American Institutes for Research
http://www.air.org/

Council of Chief State School Officers
One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
Phone (202) 336-7000
Fax (202) 408-8072
www.ccsso.org

Statement of Purpose

This report describes commonly unrecognized sources of error variance (or random variations in assessment results) and provides actions states can take to identify and reduce these errors in existing and future assessment systems. These "errors" are not mistakes in the traditional sense of the word, but reflect random variations in student assessment results.

Unlike the technical report on which it is based,[1] this report is written for policymakers and educators who are not assessments experts. It provides an explanation of the sources of error variance, their impact on the ability to make data-based decisions with confidence, and the actions states and their contractors should take as quickly as is feasible to improve the accuracy and trustworthiness of their state testing program results.

[1] CCSSO's TILSA collaborative reports on both Phillips's research and the peer review panel's response in a paper titled *Addressing Two Commonly Unrecognized Sources of Score Instability in Annual State Assessments*. The paper contains technical explanations of the two sources of error variance that Phillips found and identifies best practices that both Phillips and the expert panel recommend states adopt to minimize these sources of error variance. The paper can be found at http://www.ccsso.org/Documents/2011/Addressing%20Two%20Commonly%20Unrecognized.pdf

Introduction

State testing programs today are more extensive than ever, and their results are required to serve more purposes and high-stakes decisions than we might have imagined. Assessment results are used to hold schools, districts, and states accountable for student performance and to help guide a multitude of important decisions: Which areas should be targeted for improvement? How should resources be allocated? Which practices are most effective and therefore worthy of replication? Test results play a key role in answering these questions.

In 2014, the consequences associated with state test results will increase as new multistate assessments developed under the Race to the Top (RTTT) Program are launched. Under this program, assessment results must be appropriate for helping to determine

- student proficiency;
- student progression toward college/career readiness;
- student school-year growth;
- teacher effectiveness (teacher evaluations); and
- principal effectiveness (principal evaluations).

With so much at stake, test results must be as accurate as possible…

… if state test scores are not sufficiently accurate, they cannot help to guide the country's educational systems where they need to go.

With so much at stake, test results must be as accurate as possible. Policymakers need to trust that test scores will correctly distinguish test takers who are *at* or *above* the desired level of proficiency from those who are not. Educators need to be able to use the results to identify the effectiveness of instructional interventions and curricular changes. In short, we must be confident that any year-to-year changes in test scores—up or down—are due to *real* changes in student performance and not changes related to variation in student performance from one time to the next that get introduced during the development of the tests. These random fluctuations are referred to by testing specialists as "error variance" and are different from "mistakes." This characteristic of assessment scores cannot be eliminated, but it can be identified, minimized, and taken into account when reporting results from a testing program. If this is done, the interpretations of the scores will be much more valid.

Simply put, if state test scores are not sufficiently accurate, they cannot help us to guide the country's educational systems where they need to go.

Background

Are test scores as accurate as they should be?

Recent research by Gary Phillips of the American Institutes of Research suggests that state testing results are often less accurate than commonly believed. Measurement experts have always known that test scores have some level of uncertainty. However, Phillips investigated some unexpected and hard-to-explain changes in a testing population's scores over time and identified two commonly unrecognized sources of error variance or random variability that exist in many, if not most, state testing programs: *sampling error variance* and *equating error variance*, both of which will be explained further in this report.

*Recent research suggests that state testing results are often **less accurate** than commonly believed.*

These sources of error variance are significant due to the scope and scale of their influence. While most measurement error variance inherent in individual student scores essentially disappears when those scores are aggregated to higher and higher levels, the sources of error variance identified by Phillips persist. As such, they can have substantial impact on group level results, such as year-to-year changes in the percentage of students scoring at or above proficiency in a school, district, or state.[2]

Thus the implications of Phillips's work are both startling and dramatic: if left unaccounted for, the amount of error variance in some states' accountability test results may be great enough to result in decisions based on year-to-year changes that are not only wrong, but may have harmful consequences for educators and their programs.

[2] See page 13 for an explanation of why these sources of error variance have negligible impact on individual student scores and determinations of proficiency.

Verifying the Problem

Because of the potential significance of Phillips's investigation, the Technical Issues in Large Scale Assessment (TILSA) state collaborative of the Council of Chief State School Officers asked a select group of senior measurement experts to review Phillips's research and findings.[3] Their unanimous conclusion was that

1. the problems brought forward by Phillips are real;
2. the impact of these problems can be quite substantial; and
3. the problems warrant aggressive action by state assessment personnel and their testing contractors to minimize their impact.

Sources of Error Variance

According to Phillips, test scores are probably less precise than state officials realize because of two key practices during test development.

The first practice involves how samples of students are selected to participate in the field-testing of new test questions and the subsequent analysis of data. In most cases, the way students are selected results in complex samples—not random—but the data are subsequently analyzed *as if they were* from a simple random sample. By doing this, the testing program underestimates the sampling error variance in calculating the statistics. We refer to this as *sampling error variance.*

The second practice is the failure of states to adequately calculate and report the error variance associated with adjusting for the slight differences between the annual versions of a test. We refer to this as *equating error variance*.

One conclusion from a select group of senior measurement experts:

These problems warrant aggressive action by state assessment personnel and their testing contractors to minimize the damage.

These problematic practices result in *commonly unrecognized sources of error variance* in large-scale assessment programs—factors that cause test scores to change over time for reasons that are unrelated to the knowledge or skills that the tests aim to measure.

[3] For the names and qualifications of the members of the review panel, see Appendix A on page 16.

Sampling Error Variance

As new test questions (test items) are developed for state assessments, they must be *field-tested* and evaluated before being placed in operational forms. This is typically done by embedding small subsets of the new field-test items within different forms of the current operational test. Each of these forms contains all the items used to produce a student's score, plus a unique subset of the new items to be field-tested. Then these forms are distributed to different groups or samples of the student population. Since students cannot tell the difference between the field-test items (which will not count toward their final score) and the "real" test items, their performance on field-test items provides valuable information to test developers who use that information to determine which new items might be used for future operational tests.

Ideally, field-testing would be conducted with a simple random sample of the student population—a sample that is representative of each subgroup for which results will be reported. While the practice of selecting a sample (known as *sampling*) is well founded in scientific and statistical methods, the ways in which most states currently perform this task results in a significant underestimation of the actual magnitude of the sampling error variance.

Sampling Error Variance Explained

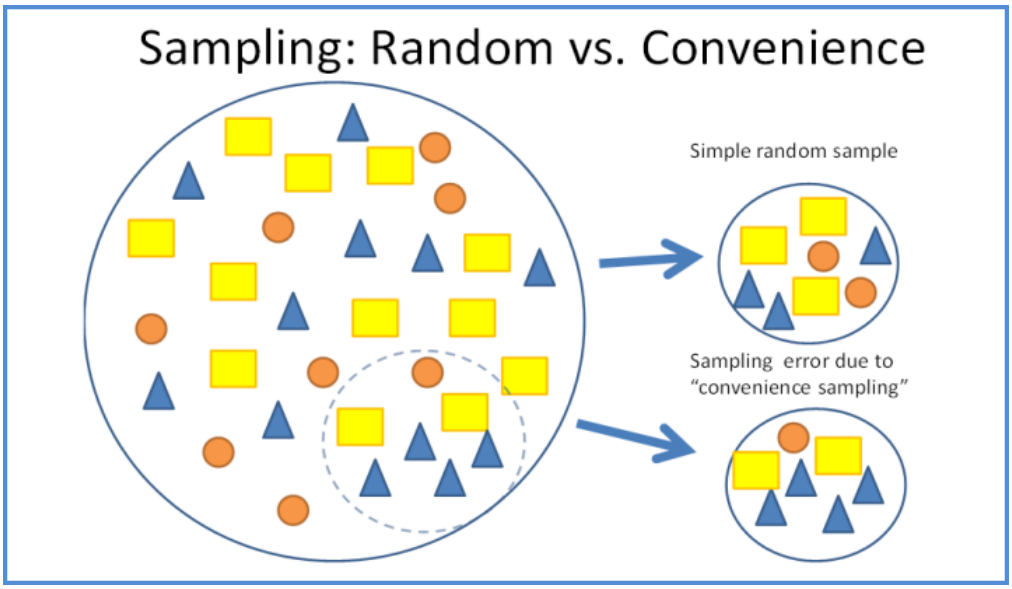
In a simple random sample, each member of the population has an equal and independent chance of being selected. Although any given sample may not perfectly represent the population, in the long run a true random selection process provides measurement experts with confidence that results from the sample will be unbiased and *generalize* to the larger group.

In states using online testing systems, it may be possible to achieve simple random samples of the student population and subgroups during field-testing.

Unfortunately, this is currently the exception—not the rule.

For states that use pencil-and-paper test booklets, true random sampling is typically not a viable choice because assigning students different test forms based on a statewide random sample, or stratified random samples from each subgroup, is logistically unmanageable. Most states therefore must use some form of *cluster sampling*, such as "convenience" samples, that rely on existing groups of students in classrooms, schools, or districts, to achieve their desired sample size. However, this practice introduces error variance because these existing groups are almost always made up of students that are more similar to each other than a group of students who are randomly

selected from the whole population. That is, the students in these clusters are not truly independent of one another and that lack of independence in the sampling can and must be taken into account.



The magnitude of this error variance will depend largely on the *spiraling* strategy selected by the state—that is, the manner by which the test forms (and therefore, field-test items) are assigned to groups (clusters) of students across the state.

A Closer Look: Sampling and Spiraling Strategies

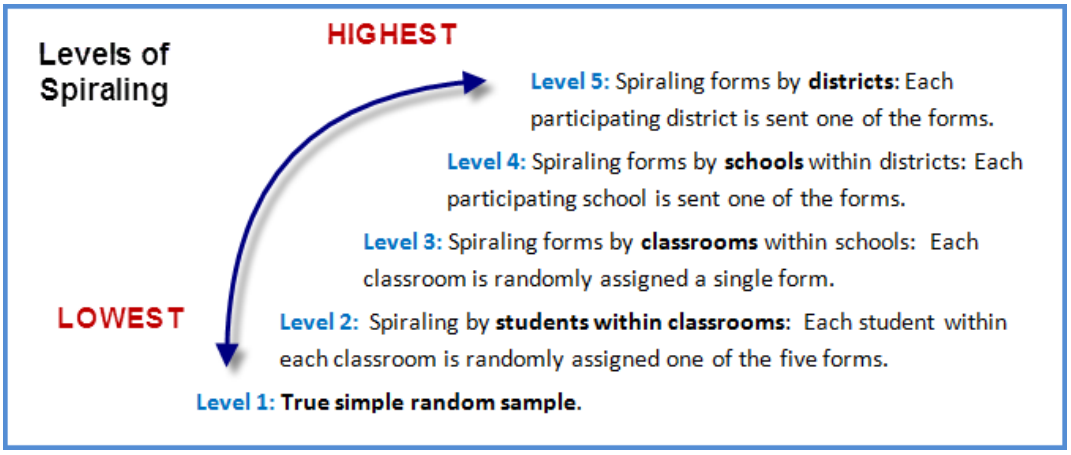
Let us use a hypothetical example to illustrate the central problem:

State A has elected to develop 30 new reading items for grade three. They know the items must be field-tested, but for a host of reasons (cost, logistics, policy, etc.) they do not want to create a separate test containing just these 30 items. The state also knows that they don't need every third-grade student to answer every new question; they just need to see how a sample of the third-grade population will perform. So the state decides to take advantage of the fact that all third-grade students will be taking the annual version of the state reading exam by creating five forms of that test: A, B, C, D, and E; that way, the state can "embed" a different subset of these 30 new items into each form (A through E) of the annual test to get the needed information.

State A further articulates the test design: if they add just six field-test items to each of the five operational forms (6 new items X 5 test forms = 30 field-test items), they will gather enough data to evaluate how suitable the new items may be for future operational tests. To fulfill the plan, state A then constructs all five

test forms in such a way that field-test items are sprinkled among, and indistinguishable from, the operational items. (This is a common procedure used in college admission, professional licensure, and other testing programs, and it ensures the testing conditions and student motivation for the field-test items match that of the operational test.)

Next, state A must determine how to spiral, or distribute, these five test forms among the third graders in the state. There are four commonly used spiraling designs in addition to a true simple random sample:



Operationally, level 5 is the easiest to implement, with the least administrative effort. Logistical concerns become progressively more complex as states move downward to level 2 (spiraling within classrooms). Dividing and distributing five unique test forms within each classroom provides a host of challenges in printing, packaging, and shipping test booklets. The state must also work diligently to avoid potential pitfalls involved in distribution, collection, and scoring the answer sheets linked to each booklet.

Typically, there is an obvious tension between operational feasibility and measurement accuracy when choosing a spiraling design.

Spiraling: Cost vs. Accuracy

Clearly, it is far more efficient and less expensive to give a single test form within each district. But from a measurement perspective, sampling error variance is greatly minimized when spiraling is done at an individual student level, such as level 1 or 2. With each upward step toward level 5, the error variance associated with sampling grows, because the sampling becomes more and more *clustered*; that is, the students who make up these "convenience" samples begin to look more similar to each other than they

would if chosen through a true simple random selection process, and the data gathered becomes less representative of the testing population on the whole.

More specifically, Phillips demonstrates in his analysis that the statistics derived from clustered samples and the quality of field-test results decreases dramatically for spiraling designs at levels 4 and 5. Therefore, he urges that states never spiral at a level higher than level 3 (the individual classroom). In addition, even for designs 2 and 3, when cluster samples are used, that clustering deceases the effective size of the sample and additional students will need to be added to the sample to assure the state's desired effective size is achieved.

> ***Clustering*** *causes the effective size of the sample to be smaller than what it appears which in turn has an impact on all item statistics.*

The Software Issue

The problem of sampling error variance is often exacerbated by the computer software states and test developers commonly use to analyze test results and produce item statistics. Why? Because the formulae underlying the calculations are often based on the assumption that true *simple* random samples were used when gathering data. (Most users are unaware of this.)

Even if a state spirals forms at the classroom level, the amount of error variance present in some of the assessment program field-test item statistics may be much larger than reported by current software. Thus the actual amount of error variance goes unrecognized. Error variance introduced at this early stage is then perpetuated and exacerbated throughout the rest of the assessment program and its subsequent stages of development. Fluctuations in aggregated scores from year to year (both up and down) may be a reflection of the imprecision inherent in the sampling procedure used—not actual improvements or declines in student performance.

What can states do about *sampling error variance*?

To reduce and manage sampling error variance, states can do the following:

1. Use true random sampling, if feasible.

States that utilize computer-based testing should investigate the feasibility and cost of implementing random sampling during field-testing. This can be done by programming the computer to administer items to a true random sample of students.

2. Use scientific sampling methods and software that account for the sampling design.

For states that cannot implement simple random samples in the field-testing of items in their current testing programs, scientific sampling methods should be adhered to and forms should be spiraled at or below the classroom level. Scientific sampling methods take into account the level of cluster sampling and the spiraling design to determine the number of students needed in the field-test to produce acceptable ranges of error variance.[4] If scientific sampling expertise is not currently available within the state department, the state's test provider or an outside consultant can provide it.

After field-testing is completed, testing personnel conduct analyses to develop a series of statistics that form the foundation of the testing program, including the calculation of equating error variance as described below. It is essential that the software package used for these analyses takes into account the specific sampling and spiraling design implemented.

[4] For a more detailed explanation, please see the CCSSO/TILSA technical report titled *Addressing Two Commonly Unrecognized Sources of Score Instability in Annual State Assessments*.

Equating Error Variance

Equating error variance is the second source of commonly unrecognized error brought forward by Phillips's work.

For even the most experienced test developers, it is nearly impossible to create different annual versions of a test in such a way that all versions are perfectly equivalent in difficulty. Still, measurement experts can employ practices to account for these differences, and *equating* is the methodology they use to make scores comparable across different versions of the test. The results of equating allow states to compare scores across different school years regardless of which versions are administered; in other words, equating makes adjustments for slight differences in difficulty between different versions of the test.

However, when states fail to properly calculate the *equating error variance*, then it may be underestimated. In many cases the software used by test developers for data analyses can lead to equating error variance that is much larger than states realize. States and contractors must use newer versions of software packages that take into account the specific sampling and spiraling design implemented. This unrecognized equating error variance impacts large group scores—such as school, district, or state average scores or proficiency rates—much more than individual or small group scores.

Phillips and the panel agreed that equating error variance is almost universally unrecognized and may be the main source of instability in state and district results.

Equating Error Variance: Case Study

Phillips, using three years of longitudinal data from three different state assessment programs, demonstrated the impact of failing to properly calculate equating error variance for each of the various annual state tests. This was an impressive set of 126 examples: each of three states assessing two subjects (reading and mathematics) across seven grade levels (three through eight, plus once in high school) over three consecutive years. For each of the 126 examples (3 x 2 x 7 x 3 = 126) Phillips started with the percentage of students scoring proficient or better on that particular test. He then estimated the equating error variance associated with each of these tests, but based the calculation on the *inappropriate assumption that the item data developed during field-testing had come from a simple random sample of students*. These flawed estimates of equating error variance were used to construct margins of error or "standard errors of a percent" for each of the 126 state scores expressed as the percentage of students

scoring proficient or better at the state level. Typically these margins of error were very small, on average slightly less than one-half of one percentage point, because they were associated with thousands of student scores. This means that if the test had been given on another day, or if a different but equivalent version had been given, one would not expect the percentage of students scoring proficient or better to be different (higher or lower) by more than one percentage point or two.

However, Phillips found that when the *correct* equating error variance was used, that is, the equating error variance that accurately reflects the effects of *cluster sampling*, the standard error of a percentage or margin of error for the state-level scores increased significantly. Instead of having confidence that if the test had been given on a different day the percentage scoring proficient or better would be no more than one or two points higher or lower, we would need to recognize that it could be possible that the proficiency rate could be off by as much as three or four percentage points higher or lower.

Why is the accuracy of the percentage proficient on the state assessment so important?

States want to know if this year's proficiency rates really are different from last year's rates. That is, are they large enough to show that the state, districts, and schools are helping more and more students to achieve proficiency in reading and mathematics. Since all assessment results contain some degree of uncertainty or error variance, measurement specialists take into consideration this uncertainty in each year's scores before claiming that scores have gone up or down. To the extent that there is a degree of uncertainty in test scores, even for large groups of students, we need to exercise caution in our inferences about the "true" results.

Although Phillips found that the margins of error were calculated incorrectly in the majority of cases, he concluded that if those margins of error had been calculated correctly, the currently reported changes in year-to-year proficiency rates typically would have been *within* the correct margin of error. This meant that the observed change that was claimed could have been nothing more than a simple reflection of random variation or error variance inherent in the scores—and therefore NOT a true change in the students' learning.

Phillips also concluded that one could only make such judgments with confidence when looking at score changes over a *two-year* period, and that in general, changes from one year to the next may be too small to extend outside the margin of error. By contrast, changes across a two-year period (such as comparisons between 2008 and 2010 results) are more likely to be large enough to extend beyond the margin of error, and therefore more likely to reflect meaningful change.

At a time when states and districts are demanding more frequent feedback to adjust and accelerate their improvement efforts, the inability to rely on year-to-year changes poses a major problem. The good news, however, is that equating error variance can be substantially reduced within large group results.

What can states do about *equating error variance*?

The panel and Phillips recommended the following actions, which must be taken in concert with the previous recommendations concerning sampling practices:

1. **Whenever possible, embed field-test items within operational tests, as opposed to running stand-alone field-tests.**

   Tests that carry no consequences can be treated quite differently by test proctors, teachers, and students than tests that carry significant consequences. Therefore states should avoid the use of stand-alone field-tests—and most states do. When this is not possible, like at the very beginning of a new testing program, it will be critical to use scientific sampling and spiral at the lowest possible level (see discussion of spiraling on page 8).

2. **Improve state procedures used to equate different versions of the state test(s) and to estimate equating error variance (as described in the technical report[5]).**

   Phillips and the panel recommended the following:

   - Items used for linking different versions of the test are selected in a manner that ensures the final set of linking items remains representative of the test blueprint.
   - Appropriate estimates of the equating error variance are always calculated, incorporated, and reported.
   - Appropriate checks on the accuracy of the equating, including replicating the equating, are incorporated when necessary.

   Policymakers may be most interested to note that Phillips and the panel recommended states take steps to ensure equating error variance is included as part of the reported confidence interval in state testing program results for each reported group, as well as in evaluations of

[5] CCSSO/TILSA technical report titled *Addressing Two Commonly Unrecognized Sources of Score Instability in Annual State Assessments*.

programs that rely on state assessment results. This will enable policymakers to more clearly distinguish meaningful change from random fluctuations.

3. **Avoid using stand-alone field-test results to establish or implement "cut scores" that distinguish student performance levels; wait until after the first operational administration to set cut scores and be prepared to revisit them after two or three years.**

Stand-alone field-test results may contain more error variance than "live" operational test results, whether due to sampling issues, lack of student motivation, or student/teacher/administrator uncertainty surrounding the introduction of new test content, new item formats, or new administration conditions. It is better to establish performance standards using data that most accurately and reliably reflect statewide student performance, even when that might result in an uncomfortable delay in releasing final test results.

Minimizing Error Variance in New Assessment Systems

The increased use of computer-delivered state assessments has the potential to considerably reduce the sampling error variance associated with field-testing items. While sampling error variance can still exist in computer-delivered tests, computerized delivery has the potential to offer much easier solutions to the sampling challenges than traditional paper-and-pencil delivery. Computer delivery allows us to do random sampling much more simply and effectively than in a paper-and-pencil environment; a computer-delivered system can be programmed to deliver the alternate test forms randomly to the entire testing population. In this way, we can be confident that each set of field-test questions has been taken by a randomly equivalent sample of students, thus eliminating the error variance associated with cluster sampling described earlier. This reduction in sampling error variance then leads naturally to much lower standard errors for the various item statistics.

SPECIAL INSERT

Why Do These Sources of Uncertainty Impact Group Results More Than They Do Individual Student Results?

Sampling and equating error variance, as discussed in this report, have a substantial and generally unrecognized impact on statewide and districtwide results. Does it then follow that these same issues impact student-level scores and determinations of individual proficiency? Not exactly.

There is no sampling error variance in an individual score, although equating error variance *does* get included in individual scores—but only to a very small extent. The scores of individual students typically vary in small ways from one occasion to another, even when there is no real change in their achievement. Testing specialists refer to this as measurement error variance, and it is very different from "mistakes." Test developers go to considerable effort to keep this error variance or uncertainty very small.

The errors introduced by sampling procedures impact test statistics, particularly statistics descriptive of test items, but not individual test scores. Equating error variance is included in individual scores; however, it is very small relative to the individual measurement error variance. When individual test scores are aggregated to produce various group scores, the resulting measurement error variance in the group score decreases dramatically because the individual measurement errors (positive and negative) tend to cancel each other out. For example, the average or mean score for a state, based on thousands of individual scores, has almost no measurement error variance. This, however, is not the case for equating error variance. Equating error variance, once calculated, is a fixed value for any set of equated tests and will be the same size no matter what the level of aggregation; so, although at the individual level equating error variance is relatively small compared to measurement error variance, as more and more scores are aggregated (and the measurement error variance gets smaller and smaller) the fixed amount of equating error variance becomes relatively much larger.

Thus the impact of sampling and equating error variance is much greater on large group scores.

Conclusion

Underestimating the instability or error variance associated with state test scores can have serious implications for our educational system. Simply put, it increases the likelihood that major decisions will be based on erroneous claims and assumptions about student performance—claims that cannot be supported with confidence when sampling and equating error variances are *properly* estimated.

In too many cases, states and districts are basing their conclusions about whether or not student achievement is improving on simple one-year-difference scores for various subgroups. These calculations then become the basis for decisions about whether to invest in curricular changes, how to target professional development initiatives, and what kinds of instructional interventions to make (changes in school programming and/or classroom practices intended to address specific academic deficiencies). Or, policymakers use test results to arrive at the decision that no such investments are needed at all.

To compound the problem, when states and districts invest in new programs, curricula, or professional development initiatives, they often rely on perceived changes in state test scores to evaluate the impact of what they have implemented. Because year-to-year declines or improvements tend to be incremental rather than dramatic, the significant amount of error variance in group scores makes it difficult to see *any* true year-to-year changes. Thus the danger is that conclusions about effectiveness will likely be based on random test score fluctuations rather than meaningful changes in achievement.

Basing weighty decisions on imprecise test scores has significant implications for public attitudes about teachers, testing, and the educational system in general.

Test scores are used for far more than instructional and curricular decisions, however. In recent years policymakers have placed increasing importance on year-to-year changes in state or district test scores: they have placed performance clauses into contracts with their superintendents, awarded bonuses, or chosen to not extend contracts based on test results. Over the long term, the Race to the Top program calls on states to use test scores for decisions on an individual educator's effectiveness.[6]

[6] As part of the grant application process, states had to agree to remove any laws that would have prevented student achievement data—such as growth in the scores of students in a given classroom or school—from being used in teacher and principal evaluations.

Over time basing weighty decisions on imprecise test scores has significant implications for public attitudes about teachers, testing, and about the educational system and policymakers in general. If states do not take steps to control and account for the common and seemingly inexplicable fluctuations in state test results, the general public may begin to question the credibility of the claims that states make on the basis of those results. With that, public confidence in the ability of education policymakers to make informed decisions and wise investments of public dollars on the basis of scores—or confidence in the usefulness of testing in general—may gradually erode, leading to either false satisfaction with the perceived increases or unjustified panic about the decreases of test scores.

Although further research is needed to more clearly understand the degree to which the practices Phillips identifies are adding to fluctuations and uncertainty in test scores, there is sufficient evidence that they play, by any definition, a substantial role, and that states should not wait to address them. Given the amount of public money that states already have been spending on state testing programs and related educational policy decisions, and given the increased weight that the federal Race to the Top program calls on test scores to carry, we urge all states and the Race to the Top Assessment consortia to take thoughtful steps to implement the recommendations outlined above.

Appendix A
Peer Review Panelists

Robert L. Brennan
Dr. Brennan is the E. F. Lindquist Chair of Measurement and Testing in the College of Education at The University of Iowa and Director of the Center for Advanced Studies in Measurement and Assessment (CASMA). He was Director of the Iowa Testing Programs at The University of Iowa from 1994–2002. Brennan authored two books on generalizability theory and co-authored a book on test equating. He has edited three other books including the fourth edition of *Educational Measurement*. He has published numerous articles in professional journals on generalizability theory, equating, scaling, performance assessment, standard setting, and domain-referenced testing. Brennan is a Past President of NCME and received the 2000 NCME Award for Career Contributions to Educational Measurement, the 2004 AERA/ACT E.F. Lindquist Award for Outstanding Achievement in Applied or Theoretical Research in the Field of Testing and Measurement and the 1997 NCME Award for Outstanding Technical or Scientific Contribution to the Field of Educational Measurement.

Steven Ferrara
Dr. Ferrara is a Principal Research Scientist at CTB/McGraw-Hill, where he is Lead Research Scientist for the District of Columbia's statewide assessments and a scientist on CTB's Standard Setting Team. Prior to joining CTB in 2008, Ferrara was a Managing Research Director at the American Institutes for Research, Director of Student Assessment for the Maryland State Department of Education, and a high school special education teacher. His research interests include cognitive demands of achievement test items; cognitive processing during standard setting; test design and achievement constructs; and assessment of students with disabilities and English-language learners. He has served on the Board of Directors of NCME and was Editor of *Educational Measurement: Issues and Practice* for the 2004–2006 volumes. He is a co-recipient of the 2006 AERA Division D award for Significant Contribution to Educational Measurement and Research Methodology.

Michael T. Kane
Dr. Kane has held the Samuel J. Messick Chair in Validity at Educational Testing Service in Princeton, New Jersey, since September 2009. He was Director of Research for the National Conference of Bar Examiners from September 2001 to August 2009. From 1991 to 2001, he was a professor of kinesiology in the School of Education at the University of Wisconsin–Madison, where he taught measurement theory and practice. Before his appointment at Wisconsin, Kane was a senior research scientist at ACT, where he supervised large-scale validity studies of licensure examinations. His main research interests are in validity theory and practice, generalizability theory, and standard setting. Kane received the 2009 Career Achievement Award from NCME.

Robert Lee Linn
Dr. Linn is a distinguished professor emeritus of education in the research and evaluation methods program at the University of Colorado at Boulder. Linn has published more than 250 journal articles and chapters in books dealing with a wide range of theoretical and applied issues in educational measurement. His research explores the uses and interpretations of educational assessments, with an emphasis on educational accountability systems. He is a Past President of both NCME and AERA, and has received numerous awards for his contributions to the field, including the ETS Award for Distinguished Service to Measurement, the E.L Thorndike Award, the E.F. Lindquist Award, the NCME Career Award, and the AERA Award for Distinguished Contributions to Educational Research.