CCSSO
Council of Chief State School Officers

ADDRESSING TWO COMMONLY UNRECOGNIZED SOURCES OF SCORE INSTABILITY IN ANNUAL STATE ASSESSMENTS

Based on research by Gary W. Phillips of

American Institutes for Research®
Making Research Relevant.

Content prepared by:

Center for K–12 Assessment & Performance Management at ETS

February 2011

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.

Addressing Two Commonly Unrecognized Sources of Score Instability in Annual State Assessments

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

Content Prepared By
Nancy A. Doorey, Center for K-12 Assessment & Performance Management at ETS
http://www.k12center.org/mission.html

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

The work reported in this paper reflects a collaborative effort of many individuals representing multiple organizations. It began during a session at the October 2008 meeting of TILSA when a representative of a member state asked the group if any of their programs had experienced unexpected fluctuations in the annual state assessment scores, and if so, whether they have a reasonable explanation for such instability in the scores. Gary Phillips, representing AIR, offered that he had been investigating what he called "score drift" and did have a hypothesis about this phenomenon and offered a brief explanation regarding what he characterized as two unrecognized sources of error that underlie such fluctuations. Dr. Phillips was invited to make a much longer presentation on his hypothesis and findings at the June 2009 meeting of TILSA. In April of 2010 Gary Phillips presented a full day pre-session at the annual meeting of the National Council on Measurement in Education (NCME) on: (1) the errors associated with the failure to use scientific sampling methodology in combination with cluster sampling designs and (2) errors associated with equating test forms. Following this session, because of the potential significance of the argument being made, the TILSA program advisors approached Dr. Phillips and ask him if he would be willing to have his hypothesis, data, and recommendations reviewed by a select group of senior measurement experts. He agreed to, and welcomed, the opportunity to have a group of his peers critique his work and a subsequent full day session was held at the CCSSO offices on September 30, 2010.

Specifically, credit goes to:

Gary W. Phillips, AIR

Members of the Technical Issues in Large-Scale Assessment (TILSA) State Collaborative (2008 to 2011)

Nancy A. Doorey, Pascale D. Forgione, and Lora Monfils, Center for K-12 Assessment & Performance Management at ETS

Duncan MacQuarrie, Robert Olsen, Douglas A. Rindone, and Charlene G. Tucker, CCSSO

Robert Brennan, Steve Ferrara, Michael Kane, and Robert Linn, NCME Expert Panel

As the national sense of urgency for the improvement of K–12 education has increased over the past several decades, state assessment data have become the primary yardstick of progress as well as the deciding factor in an increasing array of high-stakes decisions regarding both individuals and groups. The Race to the Top Assessment Program, which 46 states have joined to date, raises the stakes to new levels, requiring that the data be appropriate for use in both college entrance/placement decisions and determinations of teacher and school effectiveness. This requirement, in turn, places new demands on the accuracy of student assessment scores and the inferences drawn from them.

A panel of nationally recognized experts (see Appendix A) recently reviewed research conducted by Gary Phillips of the American Institutes for Research on potential causes of the seemingly random fluctuations seen in many annual state test scores — fluctuations that both the state agency personnel and testing vendors often find difficult to explain. Phillips has identified and investigated the impact of two underlying sources of error that contribute to the instability of these annual scores:

- Sampling error variance associated with cluster sampling of students (referred to as *design effects*) rather than true random sampling; and
- Error variance associated with equating test forms.[1]

Phillips' overarching point was that the impact of these two sources of error is substantial and is propagated throughout the assessment program statistics, eventually being seen in the unexpected instability or "bounce" of state assessment results that he refers to as "score drift." Further, he delineated the steps that can be taken within existing testing programs and within the design of new programs to minimize their impact.

The panel concluded that Phillips has identified sources of significant and often unrecognized error variance, joined him in recommending that state assessment personnel and their vendors implement specific changes to the psychometric practices that support their testing programs, and added several recommendations beyond those that Phillips had brought forward. Taking these combined actions will enable states to: (a) minimize annual score fluctuation due to error that undermine the validity of inferences from Adequate Yearly Progress (AYP) calculations, achievement growth calculations, and other portrayals of state assessment results; (b) reduce the error in student data that may be used by states or districts within evaluations of teachers, principals or superintendents; and (c) more accurately identify practices and innovations that accelerate improvements in teaching and learning.

[1] The phrase "equating error variance" as used in this document and attached slides has a different meaning from that used in most of the equating literature (see, for example, Kolen & Brennan, 2004, especially chapter 7).

Phillips points out that the error variance issues discussed herein have long been present but rarely attended to, in part because the field of educational measurement and the training programs and software to support it have historically focused on the assessment of the achievement of individuals rather than groups. As the stakes associated with group level results increase, these issues are important to address. In particular he argues, the error can be several times as large as the currently calculated estimates. This, then, results in erroneous conclusions regarding school and subgroup performance and growth. Actual margins of error can be so large as to require either very large differences in mean scores or multiple years of data to support meaningful inferences at the group level. By addressing the issues raised by Phillips, states will be able to obtain more precise estimates of error and thereby more accurately differentiate true change from random fluctuations due to error variance.

The purpose of this paper is to summarize the points of consensus reached by the expert panel regarding Phillips' work on methodological factors that contribute to score drift, and the actions state agency personnel can and should take to manage and minimize these sources of error.[2] For more details on the issues raised within this white paper, please refer to the PowerPoint presentation slides authored by Gary Phillips found in Appendix B.

Rationale and Discussion of Recommendations

As stated above, Phillips contends that there are two unrecognized sources of error variance in statistics based on state testing data that contribute to *score drift*: (a) increased error due to sampling of students, or design effects, and (b) error due to equating. Changes in psychometric practice to manage these sources of error variance can increase the ability to detect real change and draw meaningful inferences. In this section, the issues and the recommendations for increasing score stability are described in greater detail.

Sampling Error (i.e., Design Effects): Large scale assessments typically use samples of students to field test items, when conducting special studies and, depending on the program, for post-equating or linking operational forms. The most commonly used sampling methodologies in state testing programs inadvertently introduce error variance resulting in significantly greater margins of error than are recognized by either the state department or the test vendor.

The term used to describe the increase in the error variance of a statistic caused by the choice of sample design is *design effect*. The design effect is defined as the ratio of the variance of a statistic from a complex sample to the variance of the statistic from a simple random sample of the same size. Design effects differ for different subgroups (such as race/ethnicity) and different statistics. For state testing

[2] Note that the panel did not have access to mathematical proofs of results presented by Dr. Phillips, nor to empirical data that would support all conclusions.

programs, two aspects of sampling designs are involved in managing the design effect — sample selection and spiraling of forms.

Many state and district testing staff and vendors are unaware of design effects and therefore use statistical formulas that, unbeknownst to them, assume simple random sampling to calculate required sample sizes and margins of error in sample statistics. However, these states typically implement some form of cluster sampling, involving selection of intact groups from the population, such as districts, schools, or classrooms, without taking into account the reduction in the effective sample size due to clustering. If spiraling multiple forms, the effective sample size is also reduced, sometimes dramatically, when spiraling above the student level, and there is a corresponding increase in standard errors.

Design effects impact analysis for the program in multiple ways, including underestimation of required sample sizes, reduction in power, underestimation of Type I error rate[3], and increases in minimal detectable effect size (which is critical in the identification of effective and ineffective teachers, programs, schools or districts). This is caused in part by the use of prevailing software packages that assume a simple random sample. As a result, sample statistics (e.g., means), or item statistics (e.g., *p*-values, differential item functioning [DIF], and item response theory [IRT] parameters) are estimated with more error than is reported by the software.

Table 1 illustrates the impact of sample selection and spiraling design. In this example, the state wants to field test 12 test forms using a sample of 1,000 students per form. Five options for conducting the field test are shown. In Design 1, a true random sample of 1,000 students per form is drawn from the statewide enrollment. This is the ideal sample selection as it results in a design effect of 1.0 (no error introduced by sampling design) and the effective sample size is the intended 1,000 students. The next four designs assume, for purposes of illustration, a common cluster sampling design in which a random sample of 12 districts participates, each with an average of 8.33 schools, 10 classrooms per school, and 12 students per classroom.

[3] A Type I error occurs when the null hypothesis is rejected when it should have been retained. It is also referred to as a *false positive error*. For example, an intervention may appear to have had a positive impact on student achievement, when in fact there has been no improvement.

Table 1:
The Effect of Sampling Designs and Spiraling Designs on the Standard Error of Item Calibrations: Illustrated with Multi-stage Cluster Sampling using Four Types of Spiraling Designs

Spiraling Design		Design Effect	Inflation in Standard Error	Effective Sample Size
SRS Design 1	Forms are administered to a random state-wide sample	1.0	1.0	1,000
Spiraling Design 2	Forms are spiraled at the **student level** within classrooms	2.6	1.6	385
Spiraling Design 3	Forms are spiraled at the **classroom level** within schools	3.2	1.8	313
Spiraling Design 4	Forms are spiraled at the **school level** within districts	13.1	3.6	76
Spiraling Design 5	Forms are spiraled at the **district level**	22.3	4.7	45

Note: This table is an edited version of the table that appears in slide 21 in Part 1 of the Phillips presentation provided in Appendix B. The first design represents a simple random sample of 12,000 students drawn from the whole state with 1,000 students randomly administered each of 12 forms of the test. Designs 2–5 assume a multi-stage cluster sampling design where 12 forms are administered to a sample of 12,000 students in which 12 districts are sampled, within each district an average of 8.33 schools are sampled, within each school 10 classrooms are sampled, and within each classroom 12 students are sampled (12*8.33*10*12=12,000). The intra-class correlations for district, school, and classroom are .01, .10, and .15, respectively.

With this type of multi-stage cluster sample there are four ways to spiral test booklets: at the student level within classrooms, (Design 2), at the classroom level within schools (Design 3), at the school level within districts (Design 4), and at the district level (Design 5). When cluster sampling is used and forms are spiraled at the district level, as in Design 5, the resulting effective sample size is reduced from the 1,000 participating students to 45, or 4.5% of that calculated by commonly used software, and there is a nearly five-fold increase in the standard error of the item calibration. The impact is somewhat reduced when forms are spiraled at the school level within districts (Design 4), but is still substantial. There is a very significant improvement, however, when forms are spiraled at the classroom level within schools (Design 3). Even at this level, however, the assumed sample size of 1,000 has been reduced to an effective sample size of just 314, and testing program administrators will need to take the resulting error into account in subsequent statistics.

Examples of the impact of design effects on standard errors of individual item parameter estimates are provided in Table 2. Two items from an existing state item bank are shown, and each has a different design effect based on the specifics of the field test design. Item #1 has a design effect of 2.40. Because the WinSteps software assumes a design effect of 1.0 (simple random sample), it reported a standard

error of IRT parameter estimate of 0.0618. However, when a Jackknife procedure[4], which produces a truer estimate of error, was run, the estimate increased to 0.0957, or an increase of roughly 50 percent. Item #5 had a larger initial design effect, and the more accurate Jackknife analysis produced an error that was more than four times that reported by WinSteps. See slides 34 and 35 in Appendix B for more examples.

Table 2:
The Impact of Design Effects on the Error of IRT Parameter Estimates

Item	Jackknife Standard Error	WINSTEPS Standard Error	Design Effect
1	0.0957	0.0618	2.40
5	0.1324	0.0438	9.14

Note: This table is an extract of the table that appears in slide 34 in Part 1 of the Phillips presentation provided in Appendix B.

For calibration samples, this unrecognized error in the parameter estimates would carry through to any procedures using those parameter estimates including scale creation, equating to put the items or test forms on an existing scale, test assembly, creating ordered item booklets for standard setting, and obtaining student scaled scores based on the items in question (whether through adaptive or linear testing). In addition, the error in parameter estimates may cause over-identification of items for potential removal from the set of linking items used for equating and thus contribute to a failure to fully represent the content of the full test – a form of systematic error. The unrecognized error is further propagated in analyses of the student scaled scores for accountability purposes such as adequate yearly progress (AYP), growth, and value-added calculations.

In order to reduce error variance in state testing results and improve the state's ability to make meaningful inferences from changes in scores over time, it is very important, then, to minimize design effects. To do so, state assessment personnel should take action to address the following issues.

1) Issue: Cluster sampling designs reduce the effective sample size.
    a) Action: Use simple random sampling of the student population and each subgroup of interest whenever possible for all item and test analysis. This will ensure that the sample size reflects the number of independent observations in the data. When this is not possible, use scientific sampling methods, which include:
        i) Conducting a power analysis to estimate the necessary effective sample sizes and design effects (both of which impact the standard error of statistics),
        ii) Using sampling weights to ensure the sample is representative of the state as well as subgroups, and

[4] In order to obtain the Jackknife estimate, WINSTEPS was run $m - 1 = 85$ times (each time dropping a different school).

iii) Using the appropriate variance estimation methods to calculate the standard errors of all statistics.

2) Issue: Some spiraling designs reduce the effective sample size more than others.
   a) Action: When multiple forms are being used, spiral forms at the lowest level possible, ideally at the student level. Avoid spiraling above the classroom level (Design 3).
      i) For states using linear testing, whether administered on paper or by computer, there are two options:
         (1) Whenever possible, field test statistics should be obtained from simple random samples of the student population and each subgroup. Online testing, which is expanding rapidly, makes this very feasible.
         (2) When simple random sampling is not possible, spiral forms at the lowest level possible and never above the level of the classroom (Design 3).
      ii) For states using computer-administered adaptive testing, ensure that the adaptive algorithm is set to yield a simple random sample of the population as well as each student subgroup.

To summarize, use of simple random samples is recommended. If simple random sampling is not used, much larger samples will be needed to achieve the required effective sample size. Use scientific methods that take into account power analysis, the level of cluster sampling and the spiraling design to determine the number of students needed to achieve that effective sample size. Phillips recommends a minimum effective sample size of 300 for the Rasch model.

Equating Error: As noted by Phillips and the expert panel, equating error is a more pernicious — but almost universally unrecognized — component of error variance that is present in every testing program. As such, it may be the main source of instability in state and district testing results. All testing programs that equate from one administration to the next must estimate some type of equating parameter. Unfortunately, many states do not estimate the error variance in the estimate of the equating parameters. Even for those states that do estimate the equating error variance, it is almost always an underestimate because they do not take into account the design effect, which substantially increases the equating error variance. Factors contributing to error in the equating parameter estimates include measurement error variance in the old and new form, equating sample size, number and quality of items used in the equating, person error variance, and item error variance. Just as in initial field testing, design effects are not typically considered when determining the equating sample size, which may lead to unstable equating. That error is then propagated to every outcome of the testing program including conversion tables, estimates of student proficiency, AYP calculations, and minimal detectable effect size, which is critical in the identification of effective and ineffective teachers, programs, schools, or districts.

Table 3 shows one state's AYP results for 2008 and the impact of the unrecognized equating error. Column *c* shows the reported proficiency rate for each grade level, for mathematics and reading. For example, in grade 3 mathematics, the state reported that 52% of students met or exceeded the standard. Column *d* shows the standard error of the statewide proficiency rate as reported. However, this computation, which is based on simple random sampling formulae, does not include the effects of equating error on the standard error of the percent proficient. Finally, column *e* shows the standard error of equating in the percent proficient metric and is, on average, five times as large as the reported standard error. Moreover, the equating error in column *e* should be added to the reported standard error in column *d* for a more accurate estimate of the standard error of the reported proficiency rates. Equating error has a similar impact on means of scale scores, as Phillips demonstrated using actual state data across three years, and therefore this same process would be followed if Table 3 contained means of scaled scores instead of percent proficient statistics. Many schools and districts have large percentages of students scoring very near the proficiency cut score, so the implications of the large standard error are significant. The actions recommended in this paper are needed to both recognize the error and minimize it.

Table 3:
State AYP Percent Proficient in Mathematics and Reading for Grades 3–8 and 10 in 2008

a	*b*	*c*	*d*	*e*
	Subject	Percent Proficient 2008	Standard Error of Percent Proficient	Standard Error Equating
3	Math	52	.43	2.2
4	Math	48	.44	2.1
5	Math	44	.42	1.8
6	Math	42	.43	2.2
7	Math	40	.43	2.9
8	Math	35	.42	2.5
10	Math	34	.42	2.5
3	Reading	61	.42	1.9
4	Reading	61	.42	1.9
5	Reading	57	.42	1.9
6	Reading	57	.43	2.1
7	Reading	64	.42	2.8
8	Reading	66	.42	2.5
10	Reading	67	.41	2.8

Note: This table appears in slide 39 in Part 2 of the Phillips presentation provided in Appendix B.

To minimize equating error, Phillips again recommended that states should always use scientific sampling methodology in conducting all field tests and research studies. This includes:

- conducting the proper power analysis to determine sample size requirements;
- drawing random samples from a sampling frame;
- stratifying and weighting the sample;
- making adjustments for non-response; and
- using the appropriate error variance calculations for all statistics.

Secondly, Phillips recommended that states should calculate the equating error, conduct research on the components of equating error, include these analyses in technical reports and include equating error as part of all significance testing. Finally, Phillips recommended that field test spiraling designs never go above the classroom level.

The review panel supported these recommendations and stressed the importance of addressing these sources of error variance within existing and future K–12 assessment systems and research studies. The panel discussed additional issues related to these sources of error and added the following recommendations.

1) Issue: Linking items often do not fully represent the content, resulting in weak or biased linking.
    a) Actions:
        i) Ensure that the final set of linking items is large and highly representative of the content of the full-length test.
        ii) Evaluate the impact of removal of linking items on the content representation of the resulting linking set. Investigate reasons for the change in linking item performance and determine if there is a legitimate reason for removing the item from the linking set.

2) Issue: Stand-alone field testing of forms has a number of problems that lead to significant estimation error in addition to those that may be introduced by design effects. These problems include lack of student motivation, differences between the field test sample and the target testing population and, in the case of testing program start-up, opportunity-to-learn issues related to new curriculum implementation. As a result, item performance may change substantially when next administered under operational conditions.
    a) Actions:
        i) Avoid use of stand-alone field tests whenever possible. Instead, embed field test items in operational forms. If possible, use simple random sampling to distribute the field test items embedded in the operational forms. If simple random sampling is not possible, spiral the forms at the lowest level practicable and not above the classroom level. If a stand-alone field test must be used to obtain item statistics, such as at the very beginning of a new assessment, then again it will be critical to use scientific sampling to obtain the necessary effective sample size and avoid spiraling above the classroom level. Online testing programs can be designed to include assignment algorithms to produce simple random samples at both the population and subgroup levels.

ii) Don't set final cut scores based on field test results. Wait until after the first live administration when more accurate estimates of item parameters and equating error can be computed. Plan to revisit the cut scores within 2 or 3 years.

3) Issue: The context of K–12 testing has changed to include teacher evaluation, making K–12 testing more high-stakes like licensure exams; it is important to manage equating error successfully because of the impact on aggregation and equating over years.
   a) Actions:
      i) Ensure that you can articulate and validate the assumptions of the model being used (e.g., local independence [Yen, 1983], unidimensionality [Nandakumar & Stout, 1993; Stout, 1987, 1990], IRT model fit [Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991]).
      ii) When equating test forms, create two links, each of which contains highly representative linking item sets, and link back to two forms or two item banks to gain direct empirical evidence of the size of the equating error.
      iii) Run, in parallel if needed, at least two methods of equating as a check on the stability of the results across methods.
      iv) To evaluate the stability of equating over time, consider administering an old form to a sample of examinees as one element of a spiral within a given administration to see if equating results in the original conversion table.
      v) The calculation of standard errors of the statistics of interest should incorporate the appropriately estimated error variance due to equating.

Recommendations at a Glance

1. Use simple random sampling of the student population and each subgroup of interest whenever possible, for all item and test analysis. When this isn't possible, ensure that scientific sampling methodology is used.
2. When multiple forms are being used, spiral forms at the lowest level possible, ideally at the student level. Avoid spiraling above the classroom level (Design 3).
3. When selecting items for linking forms, ensure that the final set of linking items is highly representative of the content of the full-length test.
4. Avoid use of stand-alone field tests whenever possible. Instead, embed field test items in operational forms. If a stand-alone field test must be used to obtain item statistics, such as at the very beginning of a new assessment, then again it will be critical to use scientific sampling to obtain the necessary effective sample size and avoid spiraling above the classroom level (Design 3).
5. When equating test forms using a common item design, create two links, linking either to two different forms or two item banks. Conduct studies to estimate the equating error accurately.

6. Run at least two methods of equating as a check on the stability of the results across methods. Ensure that you can articulate and justify the assumptions of the model being used.
7. Avoid setting cut scores based on field test data. Wait until after the first operational administration and be prepared to revisit the cut scores after 2 or 3 years.

Conclusion

The improvement of student achievement is a national priority. Educators and policymakers are relying on assessment data to: reveal instructional program strengths and weaknesses; identify best practices for specific subgroups of students; hold schools, teachers, and students accountable; and shape strategies to accelerate the improvement of academic achievement. Doing so, however, is a fool's errand if the group change observed is within the margin of error. The purpose of this paper is to bring to light two widespread causes of error that result in erratic score fluctuations and in error margins that are significantly larger than currently calculated, very often exceeding the magnitude of the observed changes.

Further research would need to be conducted to gain a clearer understanding of degree to which the annual fluctuations in test scores, across states and designs, are due to the two methodological sources of error discussed herein, as opposed to other systematic sources of error. The panel agreed, however, that the effects of sampling and equating error are real, are likely quite substantial, and warrant action by state assessment personnel and testing companies to minimize them.

The recommended action steps are a combination of those developed by Gary Phillips and those added by a panel of four nationally recognized assessment experts who have a combined 144 years of experience in K–12 assessment and licensure and certification testing and have provided leadership to the premier professional organizations in educational measurement. They thank Gary Phillips for bringing forward these issues and urge all states to implement these recommendations, both within their current assessment and program evaluation work and within the design of future assessment systems.

References

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory. Principles and applications.* Boston, MA: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Kolen, M. J., & Brennan, R. L. (2004*). Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Measurement*, *18*, 41–68.

Phillips, G. W. (2011). Score drift: Why the results of large scale testing programs bounce around from year-to-year. (under journal review)

Stout, W. (1987). A non-parametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589–617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293–325.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187–213.

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.

Appendix B

Score Drift and Design Effects:

Why Test Results Bounce up and Down from Year to Year in Some State, District and School Testing Programs

Gary W. Phillips
Vice President & Chief Scientist
American Institutes for Research
Prepared for
CCSSO, September 30, 2010

Why do Test Results Bounce up and Down from Year to Year in Some Testing Programs?

- It does not make sense that student test results go up one year, down the next, then back up the next year.
- This presentation argues that there are two methodological reasons that may account for these unexpected and often inexplicable fluctuations

Some Frequent Reasons Given for Why Test Results Bounce up and Down

- Vender mistake
  - Wrong conversion tables
  - Incorrect equating procedures
  - scanner was not calibrated properly
  - used the wrong scoring rubrics
  - Psychometrician forgot to divide by 2
- Precipitous demographic changes
- Teachers worked hard one year but not the next
- Curriculum reform that works in the short run but not in the long run
- Change in leadership
- Cheating

There are Two Additional Reasons why Test Results Bounce up and Down from Year to Year in Some Testing Programs?

- The two reasons are that testing programs typically
  - ***Underestimate the sampling error variance*** in their statistics, and
  - ***Underestimate the equating error variance*** in their statistics.
- Consequently, there is substantially more error variance in the testing data than the testing director is aware of.
- Because the testing staff are unaware of this additional error they are
  - occasionally surprised by test result instability, and
  - not able to manage and minimize the error in the testing program.

How Does This lead to Instability in Testing Programs?

- This means that the statistics underlying the testing program (p-values, point-biserial correlations, IRT parameters, equating constants, etc.) all have a substantially larger margin of error than you think.
- This leads to two systemic problems
  - The testing director is surprised that results of the operational test do not behave as expected from the field test predictions.
  - Because the staff do not realize that there is more error than reported, they have no motivation to manage it and reduce it.

Design Effects and Equating Error

- The two culprits that cause inexplicable score drift in testing programs are
  - The impact of ***Design Effects*** in sampling students
  - The impact of ***Equating Error Variance*** in test scores

Design Effects

- What are they?
- How they affect
  - Sample sizes (*n*)
  - Power (1-β)
  - Type I error (α)
  - Effect sizes (δ)
- How they affect
  - Item parameters
  - Differential item functioning (DIF)
  - Equating error

What is a design effect?

- The *design effect* is the ratio of the error variance of a statistic taking the intra-class correlation into account to the error variance of the statistic ignoring the intra-class correlation (with the same number of cases).
- *Design effects* differ for different sub-groups and different statistics.

What is the intra-class correlation?

- The variance of student scores σ^2 has two components
  - $\sigma^2_{(between)}$ = between school variance
  - $\sigma^2_{(within)}$ = within school variance
- The intra-class correlation is
  - $\rho = \sigma^2_{(between)} / (\sigma^2_{(between)} + \sigma^2_{(within)})$
- The intra-class correlation is the proportion of variance due to clustering (in this case due to schools)

What is the intra-class correlation?

- Another definition of the intra-class correlation is that it is the correlation among observations (students) within groups (schools).

How is the intra-class correlation related to the design effect?

Let's say we estimate the population parameter μ with the sample statistic $\bar{y}$ in a sample of m schools with an average of $\bar{n}$ students. If we ignore the cluster variance, the error variance in the statistic $\bar{y}$ is

$$\hat{\sigma}^2_{\bar{y}(srs)} = \frac{\hat{\sigma}^2_{y(b)}}{m\bar{n}} + \frac{\hat{\sigma}^2_{y(w)}}{m\bar{n}} = \frac{\hat{\sigma}^2_{y}}{m\bar{n}}$$

If we take into account the intra-class correlation, then

$$\hat{\sigma}^2_{\bar{y}(cs)} = \frac{\hat{\sigma}^2_{y(b)}}{m} + \frac{\hat{\sigma}^2_{y(w)}}{m\bar{n}} = \left(1 + (\bar{n} - 1)\rho\right)\left(\hat{\sigma}^2_{\bar{y}(srs)}\right)$$

How is the intra-class correlation related to the design effect?

The design effect is

$$Deff = 1 + (\bar{n} - 1)\rho$$

How is the intra-class correlation related to the design effect?

Deff is affected by two quantities, $\bar{n}$ and ρ

When $\bar{n}$ equals 1 the *Deff* =1, regardless of ρ

When ρ equals 0.0 the *Deff* =1, regardless of $\bar{n}$

Another definition of the design effect

$$Deff = \sigma^2_{\bar{y}(cs)} / \sigma^2_{\bar{y}(srs)}$$

How does the design effect impact estimates of sample size (*n*)?

- Estimates of required sample sizes usually involve specifying
  - power (1-β)
  - alpha (α)
  - a minimally detectable effect size (δ)
- Almost all procedures to estimate sample sizes assume simple random samples and ρ =0.0 (e.g., Cohen's procedures)

How does the design effect impact estimates of sample size (*n*)?

- I will use Cohen's tables (table 2.4.1, page 52-53, 1969) to illustrate the impact of the design effect on sample size requirements.
- Lets assume we conduct a study
  - To detect a medium effect size (δ =.50)
  - Power of .80 (1-β=.80)
  - Alpha (type I error rate) α=.05, 2-tailed test

How does the design effect impact estimates of sample size (*n*)?

- According to Cohen's tables we would need *n* = 64 to meet these three criteria.
- However, if *Deff* = 4.0, then the effective sample size (*Effn*) is actually 64/4=16.
- This means we really only have 16 independent pieces of information in our sample.
- To meet the above three criteria we would need a sample of *n* = 64*4 = 256.

How does the design effect impact estimates of sample size (*n*)?

- One major process in testing programs where design effects have the most impact is in the spiraling of field test forms associated with field testing.
- Let's use a typical *sample design* and see the consequences of different *spiraling designs*.
- In the sample design the testing director wants to get 1,000 students per form for purposes of item calibration.

Sample Design: Three-stage Cluster Sampling

- Lets say a state testing director plans for the sample to be large enough so that after taking into account non-response he will have 1,000 students in each of 12 field test forms.
  - Stage 1: 12 districts
  - Stage 2: an average of 8.33 schools per district
  - Stage 3: 10 classrooms per school
  - SRS of 12 students per class
- (12)*(8.33)*(10)*(12) =12,000 students
- The intra-class correlations in the state are
  - Districts is .01
  - Schools is .10
  - Classrooms is .15

Design Effect: Three-stage Cluster Sampling

$$deff = 1+(\bar{n}-1)\rho_c+\bar{n}(\bar{c}-1)\rho_s+\bar{n}\bar{c}(\bar{s}-1)\rho_d$$

Estimating the Standard Error of Item Calibrations

The Effect of Spiraling Design of the Standard Error of Item Calibration

Spiral Design		Design Effect	Inflation in SE	Effective Sample Size
Design 1	Forms are administered to a random state-wide sample	1.0	1.0	1000
Design 2	Forms are spiraled by students within classrooms	2.6	1.6	380
Design 3	Forms are spiraled by classrooms within schools	3.2	1.8	314
Design 4	Forms are spiraled by schools within districts	13.1	3.6	76
Design 5	Forms are spiraled by districts	22.3	4.7	45

How does the design effect impact estimates of sample size (*n*)?

- The above table shows that the spiraling in the field test design has a major impact on the amount of error in the field test statistics.
- For example, some testing programs use Design 5 above and calculated standard errors as if they had a sample size of 1000 students when in fact they only have a sample size of 45 students.
- This means their statistics (e.g., p-values, point-biserials, differential item function statistics, item parameter estimates, means, correlations, equating error, etc.) have substantially more error associated with them then they are aware of.

How does the design effect impact estimates of sample size (*n*)?

- In fact Design 5 above provides a particularly dramatic example of a sampling practice in many state testing programs that results in *substantial* unrecognized error.
- Let's say you are using a three parameter IRT model and believe you need at least 1000 students to estimate item parameters (this is the industry standard).
- Many states use Design 5 above because it is administratively convenient and minimizes security risks.
- However, when you estimate item parameters and calculate standard errors you thought you had a sample size of 1000 students when in fact you only have a sample size of 45 students (the effective sample size)
- No testing director would be comfortable with only using 45 students to estimate parameters in the 3pl model but that is unknowingly what they are doing.

How does the design effect impact estimates of sample size (*n*)?

- Conclusion
  - In many testing programs estimates of sample size needed for field testing (and other procedures) are usually determined by assuming simple random samples.
  - Many testing programs therefore believe they are drawing a sample large enough to minimize errors.
  - Because they do not take into account the design effects the sample sizes are usually *gross* underestimates of what is needed.

How does the design effect impact estimates of Power (1-β)?

- Power is the probability that you will be able to detect the effect size that you are looking for in the data.
- If we calculate the power based on a design effect equal to 4.0 we find an effective sample size of 16 has an associated power of .29 (lower than the .80 we wanted)

How does the design effect impact estimates of Power (1-β)?

- Power analysis provides another dramatic example of the unrecognized impact of design effects.
- Since testing programs seldom use good sampling practices they seldom conduct accurate power analyses.
- Power analyses are usually done with procedures recommended by Jacob Cohen (Cohen, J., 1969, *Statistical Power Analysis for the Behavioral Sciences, 1st Edition, Lawrence Erlbaum* Associates, Hillsdale, 2nd Edition, 1988).
- But Cohen's procedures assume simple random sampling and will substantially overestimate the power in your study.

How does the design effect impact estimates of Power (1-β)?

- Conclusion
  - In most testing programs estimates of power are usually determined by assuming simple random samples.
  - Many testing programs therefore believe they are drawing a sample large enough to have sufficient power in their statistical analyses.
  - Because they do not take into account the design effects the power is usually underestimated.

How does the design effect impact the type I error rate (α)?

- Alpha is the probability of falsely rejecting the null hypothesis (type I error or false positive).
- If we calculate the type I error rate based on a design effect equal to 4.0 we find the type I error rate with an effective sample size of 16 is .57 (much higher than the .05 we wanted).
- In other words we thought we had .05 probability of a type I error when, in fact, we had a .57 probability.

How does the design effect impact the type I error rate (α)?

- Conclusion
  - In most testing programs estimates of the type I error rate (alpha) assumes simple random samples.
  - Many testing programs therefore believe they are drawing a sample large enough so that the type I error rate is some prescribed number (e.g. alpha = .05).
  - Because they do not take into account the design effects the type I error rate will be substantially larger than they think.

How does the design effect relate to the size of the effect you can detect (δ)?

- If we calculate the minimally detectable effect size based on a design effect equal to 4.0 with an effective sample size of 16 we find it is 1.0 (twice the size of the .50 we wanted).
- In other words, the design effect diminished our capacity to detect smaller effects.

How does the design effect relate to the size of the effect you can detect (δ)?

- Conclusion
  - In most testing programs estimates of the minimal detectable effect size assumes simple random samples.
  - Many testing programs therefore believe they are drawing a sample large enough to detect some prescribed effect size (e.g. a medium effect size equal to .50).
  - Because they do not take into account the design effects the minimum effect size they can actually detect is substantially larger.

Summary of the impact of design effects

- They reduce the number of independent pieces of information you have in the data (n)
- They reduce Power (1-β)
- They cause you to underestimate the type I error rate (α)
- They increase the minimum detectable size size (δ)

The impact of design effects on IRT item parameter estimation

- Design effects increase the error variance in estimating item parameters.
- This means the parameter estimates themselves are less accurate

The impact of Design Effects on the Error of IRT Parameter Estimates

Item	Jackknife SE	Winsteps SE	Design Effect
1	0.0957	0.0618	2.40
2	0.0949	0.0945	1.01
3	0.0565	0.0452	1.56
4	0.1063	0.0967	1.21
5	0.1324	0.0438	9.14
6	0.1353	0.0827	2.68
7	0.0631	0.0423	2.23
8	0.0877	0.0636	1.90
9	0.1383	0.0654	4.47
10	0.1637	0.0694	5.56

The impact of Design Effects on the Error of IRT Parameter Estimates

- The above table displays the first 10 of 37 items taken from a 2004 Grade 5 state reading independent field test form based on n = 1288 students randomly sampled from m = 86 schools. The average design effect was 2.73 for the 37 items.
- In order to obtain the Jackknife estimate, WINSTEPS was run m – 1 = 85 times (each time dropping a different school).

The impact of Design Effects on the Error of IRT Parameter Estimates

- Underestimating the error in IRT item parameters can lead to intractable problems in the testing program.
- Unstable Equating
  - Unrecognized instability in the item parameter estimates is directly propagated into unrecognized instability in estimates of the equating parameter estimates.
  - This instability is then propagated to every outcome of the testing program including conversion tables, estimates of student proficiency, classification consistency indices, AYP calculations.

The impact of Design Effects on the Error of IRT Parameter Estimates

- Unstable Ordered-item Booklets (OIB) in Standard Setting
  - Ordered-item Booklets are often based on calibrations obtained from independent field tests.
  - Independent field tests are usually based on *small sample sizes* with a *high degree of clustering* leading to large design effects.
  - Using these item parameter estimates to order the items in the OIB introduces unknown error in the ordering of items that panelists are relying on to set performance standards.
  - For example, an item may appear on page 23 of the OIB but it should have been on page 31.
  - This introduces unrecognized error in where the panelists set the performance standards.

The impact of Design Effects on the Error of IRT Parameter Estimates

- Unstable Conversion tables
  - Operational test conversion tables are often based on pre-equated calibrations obtained from independent field tests.
  - Independent field tests are usually based on *small sample sizes* with a *high degree of clustering* leading to large design effects.
  - Using these item parameter estimates in an operational conversion table introduces instability in the conversion table.

The Impact of Design Effects on Differential Item Functioning (DIF)

- Simple Random Sampling estimates overstates the Mantel-Haenszel (*MH*) Chi-Square statistic
- Resulting in over identifying the number of items as potentially biased
- Once the design effect is considered the DIF estimators result in fewer false-positives

The Impact of Design Effects on Differential Item Functioning

MC Items from Reading field test	*MH* chi-square		p-value of *MH* chi-square	
	SRS	With Design Effect	SRS	With Design Effect
Example Item 1				
Black v. White	40.12	23.44	.01	.06
Hispanic v. White	2.23	1.98	0.33	0.37
Male v. Female	7.09	4.75	0.05	0.12
Example Item 2				
Black v. White	2.32	2.29	0.31	0.32
Hispanic v. White	0.62	0.39	0.73	0.82
Male v. Female	5.52	5.19	0.063	0.07

The Impact of Design Effects on Differential Item Functioning

- Using Simple Random Sample assumptions when estimating DIF leads to over identifying biased items.
- This means the bias and fairness review committee will be looking at items that you said exhibited DIF when in fact they did not.
- I have found that about 20% of the items flagged for DIF using simple random sampling would not be flagged if the design effect were recognized.

Why proper Scientific Sampling is important

- We have a simulated population of

$N = 100,000$ students, $M = 1,000$ schools, $\overline{n} = 100$

$\rho = .49579,\ \mu = -.0041255,\ \sigma = 1.404415$

Why proper Scientific Sampling is important

- From this population we draw 500 bootstrap simple random samples of *n =1,000* students, and 500 bootstrap clustered samples each with $m = 50,\ \bar{n} = 20$
- Based on each sample, compute the mean to obtain 500 means (one for each sample). We can use the standard deviation of the 500 means as the measure of the SE of the sample mean.

Why proper Scientific Sampling is important

	Mean of Samples	Standard Deviation of Sample Means	Simple Random Sample SE $\frac{\sigma}{\sqrt{n}}$	Simple Random Sample Design Effect when $\frac{\sigma}{\sqrt{n}}\sqrt{1+(\bar{n}-1)\rho}$	Increase in Simple Random Sample SE $\sqrt{1+(\bar{n}-1)\rho}$
Simple Random Sample	-.005287	.0453072	.0444115	.05034409	1.133582
Clustered Sample	-.0049798	.1441515		.14336051	3.228004

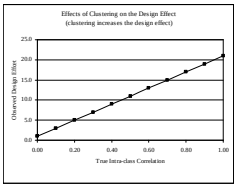
Why proper Scientific Sampling is important

- What the above slides show are:
  - The intra-class correlation is in the population of observations (in other words, the intra-class correlation is a population parameter that is estimated in the sample).
  - Any sample drawn from the population (simple random sample, clustered sample, stratified sample) will also have the same intra-class correlation among sampled observations.
  - However, whether or not the intra-class correlation inflates the *sampling distribution* of the sample statistic depends on how the sample is drawn.
  - If the sample is a simple random sample then the intra-class correlation has no effect on the sampling distribution.

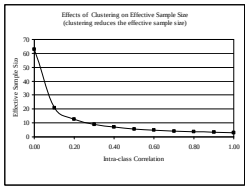
Summary of the Effects of Clustering

- Design Effects
- Effective Sample Size
- Alpha
- Power
- Minimum Detectable Effect Size
- Standard Error of The Mean

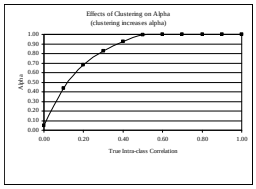
Design Effects



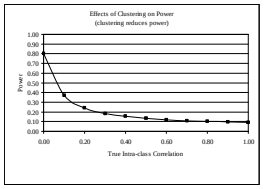
Effective Sample Size



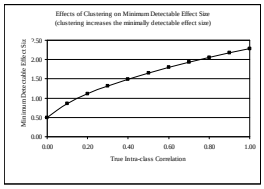
Alpha (α)



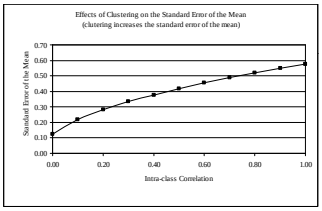
Power (1-β)



Minimum Detectable Effect Sizes (δ)



Standard Error of the Mean



Score Drift and Equating Error:

Why Test Results Bounce up and Down from Year to Year in Some State, District & School Testing Programs

Gary W. Phillips
Vice President & Chief Scientist
American Institutes for Research
Prepared for
CCSSO, September 30, 2010

Background

- In the first part of this presentation I covered the concept of design effects resulting from how you sample observations.
- The main message was that many testing programs do not know about design effects and therefore typically use the statistical formulas from simple random sampling to calculate margins of error in sample statistics.
- This practice causes the testing director to underestimate the margin of error which means their statistics can be substantially less accurate than they think they are.
- Not knowing this, the testing director confidently uses these statistics (p-values, correlations, IRT parameters, DIF statistics, etc) to assemble tests, create scales, report student results, report AYP and evaluate teachers and schools.
- Unfortunately, the unrecognized noise in the original statistics is then propagated into the scales, growth models, value-added indices and AYP calculations.
- This is like building a car where all the parts have more slippage than the engineer realizes.

Background

- In the second part of this workshop I am going to discuss equating error.
- This is an even more pernicious component of error variance that is always present in every testing program but almost universally unrecognized.
- It is probably the main source of instability in state, district and school testing results.

Equating Error Variance

- All state testing programs have to equate from one year to the next.
- The state therefore must estimate some type of equating parameter.
- Unfortunately many states do not estimate the error variance in the estimate of the equating parameters.
- Even for those states that "do" estimate the equating error variance, it is often underestimated because they do not take into account the design effect in the equating error variance nor the error covariance between items.
- In other words, even when they calculate the equating error variance, it is actually larger than they think.

What is Equating?

- Let's say we have a 2010 test in math (x = *new test*)
- We want to equate the 2010 math test to the scale of the 2009 math test (y = *old test*)
- We transform the x scale to the y scale through the equation $z = A + B\ (x)$
- Instead of reporting x we report z which is on the y scale
- The equating parameters are A (intercept) and B (slope)

Example of Equating

- Most states use Item Response Theory (IRT) common-item equating
- Although there are many other methods I will use IRT to illustrate the points. Other methods include
  - Linear equating
  - Equipercentile equating

IRT Example of Equating

- 3pl IRT Equating

$$B = \frac{\sigma_{b_y}}{\sigma_{b_x}}$$

$$A = \mu_{b_y} - B\,\mu_{b_x}$$

$$a_{z_j} = \frac{1}{B} a_{x_j}$$

$$b_{z_j} = A + B\,b_{x_j}$$

$$c_{z_j} = c_{x_j}$$

IRT Example of Equating

- Rasch Equating

$$B = 1$$

$$A = \mu_{b_y} - \mu_{b_x}$$

$$a_{z_j} = 1$$

$$b_{z_j} = A + b_{x_j}$$

$$c_{z_j} = 0.0$$

IRT Estimate of A and B

- Stocking, Lord Equating

$$SL = \sum_{q=1}^{Q}\left[\sum_{j}^{m_v} P_j\left(\theta; a_{x_j}, b_{x_j}, c_{x_j}\right) - \sum_{j}^{m_v} P_j\left(\theta; \frac{1}{B} a_{x_j}, A + B b_{x_j}, c_{x_j}\right)\right]^2$$

$$\frac{\partial SL}{\partial A} = \frac{\partial SL}{\partial B} = 0$$

What is Equating Error?

- When we estimate *A* and *B* there will be error in these estimates caused by
  - Measurement error variance in both *x* and *y*
  - Number persons used in the equating
  - Number items used in the equating
  - Person Error variance
    - Error in the estimate of parameters
    - Covariance between the errors in the estimates of parameters
    - Design effects
  - Item Error Variance

What is Equating Error Variance for Individual Scores?

$$\sigma_z^2 = B^2\sigma_x^2 + \left(\sigma_A^2 + 2(x)\sigma_{AB} + (x)^2\sigma_B^2\right)$$

for the Rasch model

$$\sigma_z^2 = \sigma_x^2 + \left(\sigma_A^2\right)$$

What is Equating Error Variance for the Mean?

$$\sigma_{\bar{z}}^2 = B^2\sigma_{\bar{x}}^2 + \left(\sigma_A^2 + 2(\bar{x})\sigma_{AB} + (\bar{x})^2\sigma_B^2\right)$$

for the Rasch model

$$\sigma_{\bar{z}}^2 = \sigma_{\bar{x}}^2 + \left(\sigma_A^2\right)$$

Methods of Estimating Equating Error Variance

- Taylor Series
  - The **delta method** is a way of deriving an approximate sampling distribution for a function of a statistical estimator from knowledge of the variance of that estimator (I use this in all my NAEP-TIMSS linking studies).
- Jackknife
  - Systematically re-compute the statistic estimate leaving out one observation (usually one primary sampling unit, e.g., a school) at a time from the sample.
- Bootstrap
  - A statistical method for estimating the sampling distribution of an estimator by re-sampling with replacement from the original sample (I will illustrate in this workshop)
- Jackknife and bootstrap yield similar results but bootstrap gives slightly different results when repeated on the same data, whereas the jackknife gives exactly the same result each time.

How to Calculate Equating Error Variance Using the Bootstrap Method

- Bootstrapping is a computer-intensive, general purpose approach to statistical inference
- It is often used as an alternative to inference based on parametric assumptions when
  - those assumptions are in doubt, or
  - require very complicated formulas for the calculation of standard errors

Demo of Bootstrapping

- Use publicly available Excel workbook (modified by me for this workshop)
- http:\\people.revoledu.com\kardi\tutorial\bootstrap\

What is Bootstrapping?

- A method for approximating the sampling distribution from the empirical distribution of the observed data.
- If we can assume the set of observations in our sample are from an independent and identically distributed population, then bootstrapping can be implemented by
  - Constructing a number of re-samples of the observed dataset
  - Each of equal size to the observed dataset
  - Each of which is obtained by *random sampling with replacement* from the original dataset

What is Bootstrapping?

- So in summary, Bootstrapping is a statistical method for estimating the *sampling distribution* of an *estimator* by *re-sampling with replacement* from the original sample
- The idea is to use a simulation, based on the actual data, to estimate the extent of sampling error

Benefits of Bootstrapping

- For large data sets, Bootstrapping and conventional estimates of sampling error usually are in agreement
- For small data sets, Bootstrapping usually provides more accurate estimates of statistical error than conventional methods
- Bootstrapping can handle almost any statistic (mean, variance, correlation, median, percentile rank, etc.)

Example of Bootstrapping the Equating Error Variance with the Rasch Model

- Two sets of 36 Rasch item parameters on two forms X (e.g., a 2010 new form) and Y (e.g., a 2009 old form)
- We wish to select a set of stable linking items that can be used to put X on the scale of Y
- We use the criteria that the difference in the linking items cannot be more than .3 logits after linking
- We use the .3 rule to iteratively determine a set of stable linking items
- To estimate the linking error we will use the bootstrap method
- We want to use 250 re-samples with replacement for the bootstrap procedure

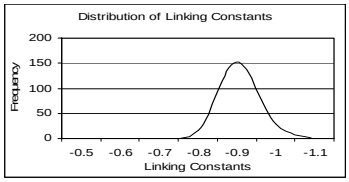
19

Example of Bootstrapping the Equating Error Variance with the Rasch Model

- The 36 item parameter estimates are assumed to be estimates with normally distributed errors from a simple random sample (this is the Winsteps assumption).
- Error variance due to sampling students: A parameter estimate is drawn from each of the 36 normal distributions 250 times.
- Error variance due to sampling items: Within each of the 250 samples, 250 non-parametric bootstrap re-samples are selected, then the .3 rule is iteratively applied.
- The total number of samples is 250*250 = 62,500, and from which, the equating SE is derived
- If there is no intra-class correlation then
  - design effect = 1

20

Example of Bootstrapping the Equating Error Variance with the Rasch Model



21

Example of Bootstrapping the Equating Error Variance with the Rasch Model

- The equating constant is -.922
- The standard error of equating is =.057
- Average of 29 items used in the equating

22

Why Large Scale Testing Programs Often Underestimate The standard Error of Equating

- **First**, many state testing programs do not even calculate the standard error of equating. Therefore, they assume equating error is equal to zero (which is never the case).
- **Second**, the industry standard is to estimate equating error due to examinee variance and not due to item variance. The variance due to items will substantially increase the standard error of equating.
- **Third**, the error in the co-variances between item parameters are almost never included in the equating error calculations. The co-variances can substantially increase the standard error of equating.
- **Fourth**, the effects of clustering in the equating sample is almost universally ignored. The design effects will also increase the standard error of equating.

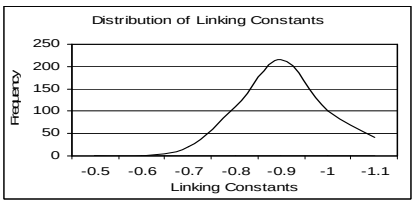
23

The Impact of Design Effects on Equating Error Variance

- The 36 item parameter estimates are assumed to be estimates with MVN distributed errors and co-variances from a complex clustered random sample.
- Error variance due to sampling students: A parameter estimate is drawn from each of the 36 normal distributions 250 times.
- Error variance due to sampling items: Within each of the 250 samples, 250 non-parametric bootstrap re-samples are selected, then the .3 rule is iteratively applied
- The total number of samples is 250 * 250 = 62,500, and from which, the equating SE is derived
- We assume there is an intra-class correlation resulting in
  - design effect = 4

24

The Impact of Design Effects on Equating Error Variance



The Impact of Design Effects on Equating Error Variance

- The equating constant is -.922
- The standard error of equating is =.107
- Average of 29 items used in the equating
- The Standard Error of Equating has been doubled by the square root of the design effect (referred to as root design effect)

Why is Equating Error Important?

Standard Error of Measurement	0.45	SE Meas. + SE Equating = 0.46	This is a 3% increase in the SE Meas.
Standard Error of the Mean	0.04	SE Mean + SE Equating = .113	This is a 283% increase in the SE Mean
Standard Error of Equating	0.107		

Note: Equating constant = -0.922, number items used in equating = 29.

Summary of Effects of Equating Error Variance

- Equating error variance has a *small effect* on the standard error of measurement
- Equating error variance has a *huge effect* on the standard error of the mean
- Equating error variance has an even larger impact on the standard error of the mean in the population than it does in a sample.
- Design effects increase the equating error variance (the standard error of equating in the above example is almost doubled by the design effect)

EXAMPLE of STATE RESULTS

Professional Disagreement on the Margin of Error in State Statistics

- Sampling statisticians do not agree on how to estimate the margin of error variance for state statistics.
- There are four schools of thought (ordered by the size of the standard error).
  1. The state is a population not a sample so there is no margin of error in state statistics because state statistics are not statistics they are parameters.
  2. Schools should be considered fixed strata so the margin of error in state statistics should be calculated with *stratified* random sampling assumptions.
  3. The particular set of students within a state should be considered a sample from a larger super-population of students so the margin of error in state statistics should be calculated with *simple* random sampling assumptions (used in most AYP reporting and is by far the most common practice world-wide).
  4. Both the students and schools within a state are considered a sample from a larger super-population of students and schools so the margin of error in state statistics should be calculated with *cluster* sampling assumptions.

Professional Disagreement on the Margin of Error in State Statistics

- What I am saying is that no matter how you calculate the sampling component of the state error variance (using any of the four methods above) that *equating* error variance should be added to the margin of error for state statistics.

State Level Mathematics and Reading for Grade 3-8 and 10 in 2007

Grade	Subject	Average Scaled Score	Standard Error Sampling	Standard Error Equating
3	M	295	0.36	1.51
4	M	294	0.33	1.24
5	M	290	0.35	1.43
6	M	288	0.29	1.50
7	M	288	0.34	2.14
8	M	276	0.33	1.68
10	M	277	0.38	2.27
3	R	305	0.35	1.26
4	R	300	0.34	1.31
5	R	302	0.29	1.10
6	R	302	0.34	1.70
7	R	306	0.32	1.73
8	R	306	0.33	1.78
10	R	308	0.36	1.89

State Level Mathematics and Reading for Grade 3-8 and 10 in 2008

Grade	Subject	Average Scaled Score	Standard Error Sampling	Standard Error Equating
3	M	299	0.34	1.77
4	M	297	0.31	1.62
5	M	293	0.33	1.38
6	M	293	0.30	1.43
7	M	293	0.31	1.83
8	M	285	0.35	2.02
10	M	281	0.37	2.17
3	R	304	0.32	1.46
4	R	304	0.34	1.51
5	R	303	0.29	1.31
6	R	305	0.32	1.48
7	R	309	0.29	1.92
8	R	311	0.31	1.71
10	R	312	0.35	2.33

State Level Mathematics and Reading for Grade 3-8 and 10 in 2009

Grade	Subject	Average Scaled Score	Standard Error Sampling	Standard Error Equating
3	M	297	0.31	1.62
4	M	300	0.34	1.67
5	M	295	0.32	1.28
6	M	293	0.31	1.40
7	M	298	0.31	2.06
8	M	290	0.38	1.98
10	M	286	0.35	1.92
3	R	308	0.33	1.32
4	R	306	0.34	1.35
5	R	308	0.32	1.81
6	R	310	0.32	2.20
7	R	314	0.33	1.63
8	R	314	0.30	1.80
10	R	317	0.33	2.01

State Level Significance Testing for Gains from 2007 to 2008

Grade	Subject	Difference 2009-2008	z-test (SRS)		z-test (SRS plus Equating Error)		Effect Size for 2008-2007	Minimum Detectable Effect Size
3	M	3.6	7.21	(*)	1.50	(ns)	0.09	0.17
4	M	3.2	3.33	(*)	1.54	(ns)	0.08	0.16
5	M	3.1	3.22	(*)	1.54	(ns)	0.08	0.14
6	M	4.4	4.69	(*)	2.10	(*)	0.13	0.17
7	M	5.0	5.18	(*)	1.76	(ns)	0.14	0.21
8	M	9.1	9.54	(*)	3.39	(*)	0.23	0.19
10	M	3.9	3.93	(*)	1.22	(ns)	0.09	0.21
3	R	-0.7	-0.71	(ns)	-0.35	(ns)	-0.02	0.15
4	R	3.8	3.90	(*)	1.83	(ns)	0.10	0.15
5	R	1.1	1.13	(ns)	0.60	(ns)	0.03	0.15
6	R	3.5	3.67	(*)	1.53	(ns)	0.09	0.17
7	R	2.7	2.89	(*)	1.03	(ns)	0.07	0.21
8	R	4.4	4.51	(*)	1.74	(ns)	0.12	0.19
10	R	3.5	3.57	(*)	1.16	(ns)	0.09	0.21

Note: Minimally detectable effect size is based on alpha = .025, power = .80 and sample size = $(n_{2007} + n_{2008})/2$

State Level Significance Testing for Gains from 2008 to 2009

Grade	Subject	Difference 2009-2008	z-test (SRS)		z-test (SRS plus Equating Error)		Effect Size	Minimum Detectable Effect Size
3	M	-1.7	-3.69	(*)	-0.69	(ns)	-0.04	0.18
4	M	3.2	6.93	(*)	1.34	(ns)	0.09	0.18
5	M	1.5	3.23	(*)	0.77	(ns)	0.04	0.14
6	M	0.6	1.42	(ns)	0.30	(ns)	0.02	0.16
7	M	5.0	11.29	(*)	1.78	(ns)	0.14	0.22
8	M	4.8	9.30	(*)	1.67	(ns)	0.12	0.19
10	M	4.8	9.45	(*)	1.64	(ns)	0.12	0.20
3	R	3.5	7.73	(*)	1.74	(ns)	0.09	0.15
4	R	2.4	4.95	(*)	1.15	(ns)	0.06	0.15
5	R	4.3	9.95	(*)	1.90	(ns)	0.12	0.18
6	R	4.9	10.75	(*)	1.83	(ns)	0.13	0.20
7	R	5.7	12.78	(*)	2.21	(*)	0.16	0.20
8	R	3.3	7.74	(*)	1.33	(ns)	0.10	0.20
10	R	4.8	9.79	(*)	1.53	(ns)	0.12	0.22

Note: Minimally detectable effect size is based on alpha = .025, power = .80 and sample size = $(n_{2008} + n_{2009})/2$

State Level Significance Testing for Gains from 2007 to 2009

Grade	Subject	Difference 2009-2007	z-test (SRS)		z-test (SRS plus Equating Error)		Effect Size	Minimum Detectable Effect Size
3	M	1.86	3.89	(*)	0.82	(ns)	0.05	0.16
4	M	6.39	13.57	(*)	2.99	(*)	0.17	0.16
5	M	4.63	9.68	(*)	2.34	(*)	0.12	0.14
6	M	5.05	11.79	(*)	2.41	(*)	0.15	0.17
7	M	10.00	21.68	(*)	3.32	(*)	0.27	0.23
8	M	13.88	27.55	(*)	5.24	(*)	0.34	0.18
10	M	8.71	16.78	(*)	2.89	(*)	0.21	0.20
3	R	2.84	5.93	(*)	1.92	(ns)	0.07	0.13
4	R	6.15	12.72	(*)	3.17	(*)	0.16	0.14
5	R	5.39	12.39	(*)	2.49	(*)	0.15	0.17
6	R	8.45	18.02	(*)	3.00	(*)	0.22	0.21
7	R	8.35	18.28	(*)	3.45	(*)	0.23	0.18
8	R	7.71	17.29	(*)	3.01	(*)	0.22	0.20
10	R	8.29	16.96	(*)	2.96	(*)	0.21	0.20

Note: Minimally detectable effect size is based on alpha = .025, power = .80 and sample size = $(n_{2007} + n_{2009})/2$

State AYP Percent Proficient Mathematics and Reading for Grade 3-8 and 10 in 2007

Grade	Subject	2007 %	Standard Error 2007	Standard Error Equating 2007
3	M	48	0.43	1.8
4	M	48	0.43	1.7
5	M	40	0.42	1.7
6	M	39	0.42	2.2
7	M	37	0.42	2.9
8	M	26	0.38	2.2
10	M	29	0.40	2.5
3	R	61	0.42	1.6
4	R	54	0.43	1.7
5	R	60	0.42	1.7
6	R	55	0.43	2.3
7	R	62	0.42	2.6
8	R	60	0.43	2.5
10	R	65	0.42	2.4

State AYP Percent Proficient Mathematics and Reading for Grade 3-8 and 10 in 2008

Grade	Subject	2008 %	Standard Error 2008	Standard Error Equating 2008
3	M	52	0.43	2.2
4	M	48	0.44	2.1
5	M	44	0.42	1.8
6	M	42	0.43	2.2
7	M	40	0.43	2.9
8	M	35	0.42	2.5
10	M	34	0.42	2.5
3	R	61	0.42	1.9
4	R	61	0.42	1.9
5	R	57	0.42	1.9
6	R	57	0.43	2.1
7	R	64	0.42	2.8
8	R	66	0.42	2.5
10	R	67	0.41	2.8

State AYP Percent Proficient Mathematics and Reading for Grade 3-8 and 10 in 2009

Grade	Subject	2009 %	Standard Error 2009	Standard Error Equating 2009
3	M	48	0.42	2.1
4	M	50	0.43	2.0
5	M	45	0.43	1.8
6	M	44	0.44	2.1
7	M	47	0.44	2.9
8	M	39	0.44	2.3
10	M	34	0.42	2.4
3	R	62	0.41	1.7
4	R	61	0.42	1.7
5	R	61	0.42	2.2
6	R	65	0.42	2.6
7	R	67	0.41	2.2
8	R	68	0.42	2.5
10	R	0.73	0.0039	0.025

State Significance Testing for AYP Gains from 2007 to 2008

Grade	Subject	Difference %	z-test		z-test using Equating Error		Effect Size for	Minimum Detectable Effect Size
3	M	4	5.83	(*)	1.22	(ns)	0.07	0.16
4	M	1	1.13	(ns)	0.25	(ns)	0.01	0.16
5	M	4	6.92	(*)	1.64	(ns)	0.08	0.14
6	M	3	4.63	(*)	0.88	(ns)	0.06	0.18
7	M	3	4.77	(*)	0.68	(ns)	0.06	0.24
8	M	9	16.10	(*)	2.72	(*)	0.20	0.20
10	M	5	8.36	(*)	1.34	(ns)	0.10	0.22
3	R	0	0.27	(ns)	0.06	(ns)	0.00	0.15
4	R	7	12.10	(*)	2.84	(*)	0.15	0.15
5	R	-4	-5.97	(*)	-1.39	(ns)	-0.07	0.15
6	R	2	2.66	(*)	0.51	(ns)	0.03	0.18
7	R	3	4.52	(*)	0.69	(ns)	0.06	0.23
8	R	5	9.00	(*)	1.48	(ns)	0.11	0.21
10	R	2	3.92	(*)	0.63	(ns)	0.05	0.22

State Significance Testing for AYP Gains from 2008 to 2009

Grade	Subject	Difference %	z-test		z-test using Equating Error		Effect Size for	Minimum Detectable Effect Size
3	M	-4	-7.23	(*)	-1.43	(ns)	-0.09	0.17
4	M	2	2.76	(*)	0.57	(ns)	0.03	0.17
5	M	1	2.38	(*)	0.56	(ns)	0.03	0.14
6	M	2	3.34	(*)	0.66	(ns)	0.04	0.17
7	M	7	11.54	(*)	1.71	(ns)	0.14	0.23
8	M	4	6.59	(*)	1.16	(ns)	0.08	0.20
10	M	0	-0.57	(ns)	-0.10	(ns)	-0.01	0.21
3	R	1	1.06	(ns)	0.24	(ns)	0.01	0.15
4	R	0	0.67	(ns)	0.15	(ns)	0.01	0.15
5	R	4	7.02	(*)	1.43	(ns)	0.09	0.17
6	R	8	12.69	(*)	2.22	(*)	0.16	0.20
7	R	2	4.18	(*)	0.67	(ns)	0.05	0.22
8	R	3	4.55	(*)	0.75	(ns)	0.06	0.21
10	R	6	10.64	(*)	1.62	(ns)	0.13	0.23

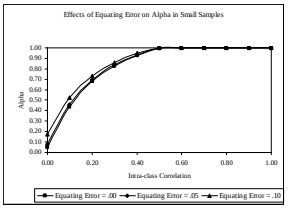
State Significance Testing for AYP Gains from 2007 to 2009

Grade	Subject	Difference %	z-test		z-test using Equating Error		Effect Size	Minimum Detectable Effect Size
3	M	-1	-1.35	(ns)	-0.29	(ns)	-0.02	0.16
4	M	2	3.89	(*)	0.89	(ns)	0.05	0.15
5	M	6	9.28	(*)	2.21	(*)	0.11	0.14
6	M	5	7.95	(*)	1.56	(ns)	0.10	0.18
7	M	0	16.32	(*)	2.39	(*)	0.20	0.24
8	M	3	22.54	(*)	4.04	(*)	0.28	0.20
10	M	5	7.78	(*)	1.28	(ns)	0.10	0.21
3	R	1	1.33	(ns)	0.32	(ns)	0.02	0.14
4	R	8	12.80	(*)	3.13	(*)	0.16	0.14
5	R	1	1.08	(ns)	0.23	(ns)	0.01	0.16
6	R	9	15.35	(*)	2.62	(*)	0.19	0.20
7	R	5	8.69	(*)	1.48	(ns)	0.11	0.20
8	R	8	13.47	(*)	2.24	(*)	0.17	0.21
10	R	8	14.58	(*)	2.41	(*)	0.18	0.21

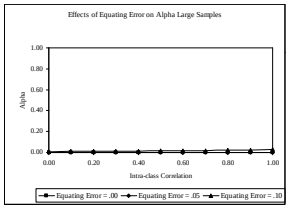
Summary of the Effects of Clustering and Equating Error

- Examine the effects of *clustering* and *equating error* in small and large samples with equal cluster sizes (21 students per school).
- Small Sample Size
  - n = 63 students; m = 3 schools; control group
  - n = 63 students; m = 3 schools; treatment group
- Large Sample Size (100 times larger)
  - n = 6,300 students; m = 300 schools; control group
  - n = 6,300 students; m = 300 schools; treatment group

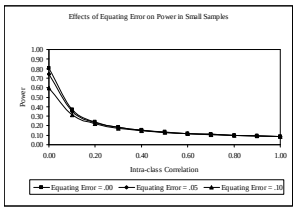
Alpha



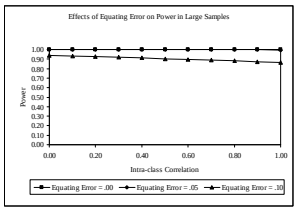
Alpha



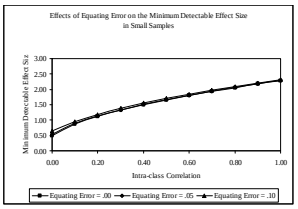
Power



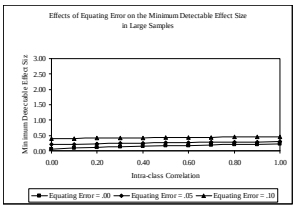
Power



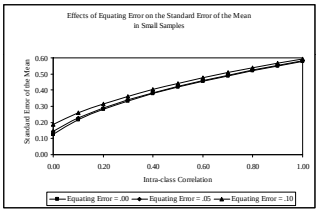
Minimum Detectable Effect Sizes



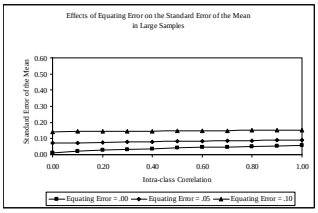
Minimum Detectable Effect Sizes



Standard Error of the Mean



Standard Error of the Mean



Summary

- In *small* samples design effects often have a big impact on the standard error of your statistics and equating error is present but has proportionally less of an impact.
- In *large* samples the design effects are still there but are often compensated for by the large sample size but equating error has a proportionally larger impact on the standard error of your statistics.

Components of the Standard Error of Equating Using the Rasch Model

Standard Error of Equating using all 44 items (MC & CR)						
Component of Equating Error	Persons[1]		Persons & Items[2]		Average Design Effect	
Covariances	Without	With	Without	With	Without	With
SRS with & without Covariance	0.013	0.038	0.055	0.066	1.0	1.0
CS with & without Covariance	0.018	0.051	0.058	0.076	2.0	2.0

[1]250 Monte Carlo simulations from normal distributions (250 equating estimates).

[2]250 Monte Carlo simulations from MVN distributions, 250 Bootstrap samples (250 X 250 = 62,500 equating estimates).

Components of the Standard Error of Equating Using the Rasch Model

Standard Error of Equating using .30 Criterion with 31items (MC & CR)						
Component of Equating Error	Persons[1]		Persons & Items[2]		Average Design Effect	
Covariances	Without	With	Without	With	Without	With
SRS with & without Covariance	0.030	0.042	0.070	0.074	1.0	1.0
CS with & without Covariance	0.042	0.056	0.077	0.086	2.0	2.0

[1]250 Monte Carlo simulations from normal distributions (250 equating estimates).

[2]250 Monte Carlo simulations from MVN distributions, 250 Bootstrap samples (250 X 250 = 62,500 equating estimates).

Components of the Standard Error of Equating Using the Rasch Model

Standard Error of Equating using .20 Criterion with 25 items (MC & CR)						
Component of Equating Error	Persons[1]		Persons & Items[2]		Average Design Effect	
Covariances	Without	With	Without	With	Without	With
SRS with & without Covariance	0.053	0.055	0.090	0.097	1.0	1.0
CS with & without Covariance	0.060	0.062	0.096	0.106	2.0	2.0

[1]250 Monte Carlo simulations from normal distributions (250 equating estimates).

[2]250 Monte Carlo simulations from MVN distributions, 250 Bootstrap samples (250 X 250 = 62,500 equating estimates).

Components of the Standard Error of Equating Using the Rasch Model

- The first component of equating error is person variance.
  - It has two sub-components:
    - variance of item-parameter estimates (employed by users of WINSTEPS, assumes SRS)
    - the covariance between item-parameter estimates
  - This is captured by the Monte Carlo simulations and is made larger by sampling design effects.
- The second component of equating error is item variance.
  - This is the variance in the difference between item parameter estimates across equating forms.
  - This is captured by Bootstrapping and is made larger by position effects, item parameter drift, motivational differences between field testing and operational testing, etc.