**The Behavior of Linking Items in Test Equating**

CSE Report 630

Edward H. Haertel
CRESST/Stanford University

May, 2004

Center for the Study of Evaluation (CSE)
National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.5 Differential Prediction and Opportunity to Learn
Strand 1: The Behavior of Linking Items in Test Equating
Edward Haertel, Project Director, CRESST/Stanford University School of Education



The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences (IES), U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences, or the U.S. Department of Education.

# THE BEHAVIOR OF LINKING ITEMS IN TEST EQUATING

**Edward H. Haertel, Project Director**

**Stanford University**

## Abstract

Large-scale testing programs often require multiple forms to maintain test security over time or to enable the measurement of change without repeating the identical questions. The comparability of scores across forms is consequential: Students are admitted to colleges based on their test scores, and the meaning of a given scale score one year should be the same as for the previous year. Agencies set scale-score cut points defining passing levels for professional certification, and fairness requires that these standards be held constant over time. Large-scale evaluations or comparisons of educational programs may require pretest and posttest scale scores in a common metric. In short, to allow interchangeable use of alternate forms of tests built to the same content and statistical specifications, scores based on different sets of items must often be placed on a common scale – a process called test equating (AERA, APA, NCME, 1999).

Formally, test equating is the process of deriving a function mapping score on an alternate form of a test onto the scale of the anchor form, such that after equating, any given scale score has the same meaning regardless of which test form was administered. Various data collection designs and analytical procedures have been used for test equating, but in the context of large-scale assessments, perhaps the most common method for year-to-year equating relies on incorporating some items from one or more previous years' tests into each successive annual test form.[1] Because these *common items* are embedded within the respective forms, they are referred to as internal linking items. This is an example of the "common-item nonequivalent groups design" (Kolen and Brennan, 1995). The embedded common items (also referred to as *anchor items* or *linking items*) provide the statistical means for equating successive test forms so that scaled scores are directly comparable.

[1] With slight modifications, the same internal anchor common-item nonequivalent groups equating design can be used for state assessments that employ matrix sampling, in which case "test form" would actually refer to a collection of linked forms constructed from the item pool for a given year.

Score scale conversions are derived from performance on these linking items, relying on the assumption that the statistical properties of each linking item are unchanged from one year to the next. If the statistical properties of items are unchanged, then any systematic difference in the proportions of examinees responding correctly from one year to the next may be attributed to differences in the ability distributions for the first year's examinees versus the second year's. This equating application has special salience in the light of P.L. 107-110, the No Child Left Behind Act of 2001, which attaches significant consequences to measured annual progress in reading and mathematics at the state, district, and school levels. Accurate linkage of successive annual forms of state tests is essential to accomplish the intent of the legislation, including implementation of the "safe harbor" provision.

This project investigated statistical problems in such year-to-year linking of state assessment data. The project involved both the development of new statistical procedures and analyses of test equating using student-level data from successive cohorts tested in each of three states.

## Findings and Conclusions

Test equating using the common item nonequivalent groups design relies on one or another model from Item Response Theory (IRT). If all of the (linking and non-linking) items in both test forms fit the IRT model, then in theory, the equating function should be invariant under different linking item selections, except for the (quantifiable) effects of random error in examinees' item responses. Year-to-year changes in performance on all linking items should follow a definite pattern, determined by the year-to-year change in the examinee proficiency distribution. In practice, however, the IRT model is only an approximation. It is not uncommon for a few linking items to show anomalous changes from one year to the next. If a linking item shows a year-to-year change in correct-response proportion that does not conform to the pattern established by the rest of the linking items, it is not used for equating. Such items are simply treated as different items in Year 2 versus Year 1, with two separate sets of item parameters.

Anomalous linking item performance might occur for various reasons. If teachers change their content emphasis from one year to the next, for example, then

items testing content elements receiving greater versus lesser emphasis may change in relative difficulty. The implementation of innovative programs, policies, and reforms in the curriculum is intended to encourage a reallocation of instructional time and resources; teachers are supposed to teach some different things from those they taught before. Thus, if an accountability system is in fact successful in encouraging a reallocation of instructional resources, then it might be that those linking items that appear anomalous from one administration to the next are precisely the ones revealing real reform effects. In theory, such differential changes might be modeled by some more complex, multidimensional IRT model, but they represent violations of the assumptions of the unidimensional IRT models actually used. (See Koretz, McCaffrey, & Hamilton, 2001, for a detailed analysis of this problem.)

During the first year of this project, we investigated whether the standard statistical procedure of setting aside linking items showing anomalous year-to-year changes might seriously underestimate true achievement gains and give a misleading picture of the effectiveness of measurement-driven reforms. When such items were removed on statistical grounds, the effects of reforms might have been adjusted away. We concluded that this scenario was highly unlikely. A detailed examination of linking item behavior was conducted at a testing company, including a review across several states of all cases where linking items had been set aside and not used. Although some linking items were found to behave anomalously, these cases almost always had obvious, uninteresting explanations (e.g., year-to-year differences in item formatting, very small item revisions, or changes in item context effects due to differences in the linking item's position in the Year-1 vs. Year-2 tests). Unexplained cases of anomalous linking item performance were far too rare to enable any systematic exploration of the relation between linking item behavior and alignment with state standards.

However, these explorations did serve to document the substantial magnitude of anchor item selection effects on equating transformations. About halfway through the first year of the project, we turned from our initial "fixed effects" focus (examining properties of specific items) to a random effects treatment of the anchor item selection process. We have demonstrated that sampling of anchor items is a significant, largely unrecognized source of statistical error in test equating.

Equating transformations are derived from empirical observations, and therefore subject to statistical error. Error may arise both from the sampling of examinees and from the sampling of items used to estimate the equating transformation. At the level of individual students' scores, total equating error from both these sources is typically negligible. However, equating transformations are used not only to place individual examinees' scores from different test forms onto a common scale, but also to place means and other distributional summaries (e.g., percent "proficient") for large groups of examinees onto a common scale. The standard error of a group mean is much smaller than the standard error of a score for an individual examinee, by a factor on the order of $1/\sqrt{N}$ , where $N$ represents the size of the group.[2] Thus, the effects of linking items' departures from IRT model assumptions loom much larger relative to the much smaller standard errors of aggregate-level statistics.

Current standard practice in test equating ignores the effects of IRT model misspecification. The standard error of the equating function is calculated treating common items as fixed (i.e., assuming that the unique characteristics of each item are fully accounted for by its item parameters). Under this assumption, the only source of imprecision in equating functions arises from uncertainty in the estimation of the item parameters (e.g., see Kolen & Brennan, 1995). Because item parameter estimates are typically based on very large samples of examinees, this uncertainty is quite small. Thus, standard practice typically gives a very reassuring picture of the precision with which equating transformations can be determined. The equating transformation is determined by the responses of thousands of examinees; the concern of this research is that these thousands of examinees are responding to, at most, dozens of common items.

In order to quantify the uncertainty in equating transformations due to common-item selection, we developed both bootstrap and analytical procedures for estimating an error component representing the random selection of anchor items

[2] Note that the standard error of a group mean has two components. One component arises from the sampling of the group tested from some (actual or hypothetical) population. The other arises from uncertainty in the estimation of each individual examinee's score. The expected value of the standard error of a randomly chosen individual examinee's score divided by $\sqrt{N}$ would represent only the second of these two components. A more detailed treatment is not required for purposes of this exposition.

from a hypothetical pool of possible items. For equating by the mean/sigma method, a formula for quantifying the standard error due to the sampling of the common items was derived using the "delta method" (Stuart & Ord, 1994). The analytic formula relies on the assumption of bivariate normality of the IRT difficulty parameter estimates. The derived standard error and a bootstrap approximation for the same quantity were calculated for a statewide assessment under both one- and three-parameter logistic IRT models; for polytomous items, a graded response model was fitted. For the one-parameter logistic case, a small-sample bootstrap approximation to the standard error of equating due to the sampling of examinees was derived for comparison purposes.

There was some discrepancy between the analytic and the bootstrap approximation of the error due to the sampling of common items because the difficulty parameter estimates did not meet the assumption of bivariate normality. To confirm the accuracy of the analytic formula, an additional "analytical bootstrap" was conducted. For simulated data drawn from a population that was distributed as bivariate normal, the two methods for estimating the error gave nearly identical results, confirming the correctness of the analytic approximation.

Illustrative findings are shown in Table 1, showing findings for a representative state assessment. As shown, equating error components due to common item sampling (column 1) and due to examinee sampling (column 2) were of about the same magnitude. In other words, the conventional standard error of the equating function reflects only about half the equating error variation. For individual examinee scores (top half of table), these two equating error components together comprised only a small proportion of the total error variance; measurement error was the largest component in individual score variability. (Note that relative contributions vary as a function of examinee ability, $\theta$.) For the state mean score, though, the picture was quite different. Measurement error in score summaries shrinks as sample size increases. Examinee-sampling equating error also decreases as samples become larger. Error due to common-item sampling does not depend on the size of the examinee sample—it is affected by the number of common items used—so it comes to constitute the dominant source of error for summary scores as N increases.

Table 1

Relative Size of Errors for Individual and Group-Level Score Interpretations for the Mathematics 8 Assessment Under a 1PL IRT Calibration

| $\hat{\theta}$ | Source of Standard Error (Percentage of the total variance[a]) | | | Total Variance |
|---|---|---|---|---|
| | Common-item sampling | Examinee sampling | SE($\hat{\theta}$) | |
| For interpreting individual scores | | | | |
| -2 | 0.05076 (2.2%) | 0.07470 (4.9%) | 0.32654 (92.9%) | 0.11479 |
| -1 | 0.03264 (3.2%) | 0.03669 (4.0%) | 0.17620 (92.8%) | 0.03346 |
| 0 | 0.02931 (2.8%) | 0.01093 (0.4%) | 0.17369 (96.9%) | 0.03115 |
| 1 | 0.04424 (3.4%) | 0.04357 (3.3%) | 0.23265 (93.3%) | 0.05798 |
| 2 | 0.06603 (3.6%) | 0.08176 (5.5%) | 0.33137 (90.9%) | 0.12085 |
| For interpreting mean scores | | | | |
| | Common-item sampling | Examinee sampling | $\sqrt{Var(\hat{\theta})/N}$ | |
| 0 | 0.02931 (82.6%) | 0.01093 (11.5%) | 0.00787 (6.0%) | 0.00104 |

[a]Percentages may not add to 100% because of rounding

## Summary

We set out to investigate systematic differences in anchor (linking) item behavior as a function of alignment with state standards, in the context of high-stakes reform. Although some linking items were found to behave anomalously, these cases almost always had obvious, uninteresting explanations (e.g., year-to-year differences in item formatting, very small item revisions, or unequal item context effects when the anchor item appeared in the Year 1 vs. Year 2 tests). However, these explorations did serve to document the substantial magnitude of anchor item selection effects on

equating transformations. Our investigation of these effects led us to consider anchor item selection effects more generally. We developed both bootstrap and analytical procedures for estimating an error component representing the random selection of anchor items from a hypothetical pool of such items, and showed that common item sampling constitutes a major overlooked source of error in test equating.

**Work Products and Dissemination**


Haertel, E. H., & Michaelides, M. P. (2003, March). *The Behavior of Linking Items in Test Equating: Interim Progress Report*, January 1, 2002 - December 31, 2002. CRESST Report.

Michaelides, M. P. (2003). *Effects of common-item selection on the accuracy of item response theory test equating with nonequivalent groups*. Unpublished doctoral dissertation, Stanford University, School of Education.

Michaelides, M. P. (2003). *Sensitivity of IRT equating on the behavior of test equating items*. Paper presented at the meeting of the American Educational Research Association, Chicago, Illinois.

Michaelides, M. P., & Haertel, E. H. (2004, April). *An Application Of A Mantel-Haenszel Procedure To Identify Misbehaving Common Items In Test Equating*. Paper accepted for presentation at the meeting of the American Educational Research Association, San Diego, California.

Michaelides, M. P., & Haertel, E. H. (2004, February). *Sampling of Common Items: An Unrecognized Source of Error in Test Equating* (Draft). CSE Technical Report submitted for CRESST dissemination.

## REFERENCES

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

Kolen, M. J., and Brennan, R. L. (1995). *Test Equating Methods and Practices.* New York: Springer.

Koretz, D. M., McCaffrey, D. F., & Hamilton, L. S. (2001). *Toward a framework for validating gains under high-stakes conditions* (CSE Technical Report 551). Los Angeles, CA: CRESST/CSE, University of California, Los Angeles, Graduate School of Education and Information Studies.

Stuart, A., & Ord, J. K. (Eds.) (1994). *Kendall's Advanced Theory of Statistics* (6th ed.). New York: Halsted Press.