Sampling of Common Items:
An Unrecognized Source of Error in Test Equating

CSE Report 636

Michalis P. Michaelides
The College Board

Edward H. Haertel
CRESST/Stanford University

July 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
GSE&IS Building, Box 951522
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

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002-03, 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.

SAMPLING OF COMMON ITEMS: AN UNRECOGNIZED SOURCE OF ERROR IN TEST EQUATING[1]

Michalis P. Michaelides
The College Board

Edward H. Haertel
CRESST/Stanford University

Abstract

There is variability in the estimation of an equating transformation because common-item parameters are obtained from responses of samples of examinees. The most commonly used standard error of equating quantifies this source of sampling error, which decreases as the sample size of examinees used to derive the transformation increases. In a similar way of reasoning, the common items that are embedded in test forms are also sampled from a larger pool of items that could potentially serve as common items. Thus, there is additional error variance due to the sampling of common items. Currently, common items are treated as fixed; the conventional standard error of equating captures only the variance due to the sampling of examinees.

In this study, a formula for quantifying the standard error due to the sampling of the common items is derived using the delta method and assuming that equating is carried out with the mean/sigma method. 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 are calculated for a statewide assessment under both three- and one-parameter logistic IRT models; for the polytomous items, a graded response model is fitted. For the one-parameter logistic case, a small-sample bootstrap approximation to the standard error of equating due to the sampling of examinees is 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. Examination of the assumption of bivariate normality of the difficulty parameter estimates showed that the assumption does not hold for the data set analyzed. 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. The comparison with the examinee-sampling standard error of equating revealed that the two sources of equating error 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. Numerical results demonstrate that for individual examinee scores the two equating errors comprised only a small proportion of the total error variance; measurement error was the largest component in individual score variability. For group-level scores though, the picture was 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

[1] We would like to thank Michael Nering and Kevin Sweeney of Measured Progress Inc. for providing the data analyzed in this study and David Rogosa and Robert Tibshirani for their insightful comments.

affected by the number of common items used—so it could constitute the dominant source of error for summary scores. The random selection of common items should be acknowledged in the analysis of a test and the arising error variance calculated for proper reporting of score accuracy.

Introduction

Tests are often administered over multiple occasions, at different times and places. Testing programs do not typically administer the same form of a test on all occasions. Repeated administration of the same items would result in overexposure of the test content and jeopardize the security of the test. Examinees taking the test later would be at an advantage over those who had taken it earlier. Moreover, the item pool needs to be refreshed because some items become obsolete over time, while others become more relevant (Goldstein, 1983), and because agencies are often required to disclose some or even all items from the test after they have been used. Even without administering the same test items, it is desirable to maintain comparable information longitudinally and be able to measure change across administrations.

When examinees take alternate forms, it does not follow that the scores they earn are comparable. Even if those forms are carefully constructed to the same content and statistical specifications, they differ in statistical properties such as their degree of difficulty. Unless scores are adjusted to take account of these differences, comparisons are not fair to all examinees tested. “Only when tests are equated can it be fair to give them to different people and treat the scores as if based on the same test” (Holland & Rubin, 1982, p. 1). Equating is the statistical process that establishes comparability between alternate forms of tests built to the same content and statistical specifications by placing scores on a common scale, thus allowing interchangeable use of scores on these forms (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Kolen & Brennan, 1995). After equating, it should be a matter of indifference to examinees which test form to take (Lord, 1980).

There are various designs that can be used for test equating. Deciding on which one to implement involves issues of practicality and statistical adequacy. In the frequently used *common-item nonequivalent groups design* (Kolen & Brennan, 1995) a subset of the items, which are referred to as common, equating, linking, or anchor items, is embedded in both forms, thus creating an overlap in their content (Wainer, 1999). Those common items are used to generate the equating relationship between the two groups, without assuming group equivalence, by comparing their performance on the common items. Given that the statistical properties of the common items do not change across administrations, any systematic pattern in their difficulty indices may be attributed to differences in the average proficiencies of the two groups.

The common-item nonequivalent groups design arises in many practical situations, such as the administration and equating of test forms over successive years. The design permits the linking of longitudinal information by establishing a common metric while at the same time avoiding the problem of test overexposure and allowing the renewal of the item pool by replacing old items with new ones.

The Standard Error of Equating

As with any statistical estimation procedure, the accuracy of estimated equating relationships is of interest. Test scores are often associated with high-stakes decisions. Therefore, one should always guard against possible sources of error, attempt to quantify the amount of inaccuracy, and report it with each estimate.

The error could be *systematic* or *random* (cf. Kolen, 1988; Kolen & Brennan, 1995). Systematic error is independent of sample size and arises in different situations that fail to adhere to guidelines of a proper application: when the method of estimating the equating function introduces bias, when the assumptions underlying the methods or the models used are violated, when the equating designs are incorrectly implemented, or when the groups taking alternate forms differ substantially (Kolen & Brennan).

Random error arises because information is collected only for a sample of a population. When there is information for the whole population, the true equating function can be calculated. When only a random sample is available, the estimate of that function fails to capture the true state by some amount of error.

Standard errors of equating quantify random error, the uncertainty arising due to the sampling of examinees. As sample sizes increase, examinee-sampling error decreases. Kolen and Brennan (1995) described in detail methods to estimate the standard error of equating, which “is conceived as the standard deviation of equated scores over hypothetical replications of an equating procedure in samples from a population or populations of examinees” (p. 211). For the common-item nonequivalent groups design in particular, random groups from each of the two populations would be drawn, and the equating relationship calculated. The equating function applied on the score scale of the second group before equating would give the equivalent score scale. The standard deviation of equated scores from many resampled draws would provide an estimate of the standard error of equating.

This standard error is different at each point on the score scale. It can be summarized to an aggregate value, the *mean standard error of equating* (Kolen & Brennan, 1995) by weighting the equating error variance at each point x_i by its density $f(x_i)$ and then summing over all score points

$$\sqrt{\sum_i f(x_i) SE^2[\hat{e}_Y(x_i)]}$$

where $\hat{e}_Y(x_i)$ is the equivalent score of x_i on a scale Y.

It is useful to have closed analytic formulas or large sample approximations for standard errors. The delta method has been used to derive such expressions for various equating designs and methods, with the exception of item response theory (IRT) methods (cf. Kolen & Brennan, 1995, Table 7.2, for a list of references for specific designs and methods.) These formulas are often complicated. Resampling procedures, such as the bootstrap (Efron, 1982; Efron & Tibshirani, 1993), even though computationally more intensive, offer feasible alternative ways for estimating standard errors of equating. In practice, standard errors of equating are often not computed because simple analytical formulas do not exist, and bootstrap or jackknife computations are impractical to implement (Harris & Crouse, 1993).

Until recently, there have not been any analytic procedures to estimate standard errors of equating for IRT equating methods. Lord (1982) derived a formula for the asymptotic standard error of a true score IRT equating with an external anchor test, noting however that his approach underestimated the error. In a series of recent articles, Ogasawara (2000, 2001a, 2001b, 2001c) provided analytic expressions for IRT equating methods. He derived asymptotic standard errors of equating coefficient estimates obtained by moments (e.g., for mean/sigma and mean/mean methods; Ogasawara, 2000), asymptotic standard errors of the estimates of equating coefficients using response functions (i.e., Haebara or Stocking and Lord characteristic curve methods; Ogasawara, 2001c) and asymptotic standard errors for IRT true score equatings (Ogasawara, 2001a).

With one exception, no implementations of computational procedures have appeared in the literature. Although Kolen and Brennan (1995) mention that bootstrap standard errors of equating can be used for IRT methods as with other methods, they note that such a task is difficult since in addition to drawing random samples, item parameters must be estimated many times. Item parameter estimation requires use of specialized software that often requires some manual intervention (e.g., dropping items or examinees; constraining particular parameters) in order to obtain a convergent solution, especially when a three-parameter IRT model is employed. For that reason, the task of estimating parameters for many resamplings of examinees becomes overwhelming.

One recent study calculated bootstrap standard errors of IRT equating methods for the common-item nonequivalent groups design (Tsai, 1998; Tsai, Hanson, Kolen, & Forsyth, 2001). That study involved bootstrap sampling from two nonequivalent samples of examinees, and for each of 500 replications they calibrated their dichotomous data using BILOG 3 to obtain item parameters. In fact, the authors mentioned that in 99 of the 500 replications for one of the conditions they examined, at least one of the IRT b estimates diverged toward $\pm\infty$, and they had to replace those estimates with values from the original calibration.

The Selection of Common Items as an Additional Source of Variability in Equating

A test score is based on an examinee's performance on a particular test form consisting of certain items. What is of interest most often is not how well the examinee did on those particular items at that particular occasion. Rather it is the inference drawn from that instance of performance to what the examinee could do across many other instances requiring the application of the same skills and knowledge. "The operational assessment is a sample of performance. Those who wish to apply a standard are concerned with the pupil's level of proficiency across a domain of knowledge and skills, not with the sample as such" (Cronbach, Linn, Brennan, & Haertel, 1997, p. 376).

Generalizing from a sample to a population is essentially an error-prone process. Much as in classical statistics where generalization from a sample to a population is a central concern, inference from a particular test to a universe of tests covering the same domain is another important issue for educational and psychological measurement (Brennan & Kolen, 1987).

In the common-item nonequivalent groups design for test equating, a source of error that has been ignored in the error calculation process is the selection of the set of items used as common. Current practice treats common items as fixed. Under IRT model assumptions, all of the relevant particularities of each item are accounted for by its item parameters. Thus, if IRT assumptions held and item parameters could be estimated with perfect accuracy, then the same equating function would be obtained regardless of which common items were used. Common items are, however, chosen from a hypothetically infinite pool of potential items that could appear in two or more forms and provide an anchor for equating test forms.[2] Many other sets of items that adhere to the same specifications could function as common items. Had a different collection of common items been used to perform equating, the equating transformation would differ due to inevitable small departures from IRT model assumptions. IRT models do not hold perfectly, and as a result, item parameter estimates vary across administrations because of model misspecification, as well as (examinee) sampling variability. Hence, each set would result in a different equating transformation, even with infinitely large samples of examinees. The argument just described demonstrates why the sampling of common items constitutes a source of error in test equating that needs to be quantified and reported. This source of error could be thought of as random, because it arises due to the sampling of interchangeable common items from the hypothetical population of such items. However, under a perfect model fit, the sampling of common items would not add variance to the equating transformation because item-specific properties would be fully accounted for by the item's IRT parameters; the variability arises due to model misspecification.

Current equating practice does not take into account this uncertainty in the estimation of item parameters when deriving individual or group-level scores, acting as if the fitted IRT models held perfectly. Pellegrino, Jones, and Mitchell (1999) in their report on the National Assessment of Educational Progress (NAEP) noted that ignoring the uncertainty in (both common and non-common) item parameter estimates resulted in standard errors being underestimated. They cited a need for investigation of the accuracy of the standard errors.

[2] An infinitely large pool of common items is a theoretical construct, rather than a plausible practice; items are not randomly sampled from this pool because the collection of common items should adhere to certain content and statistical specifications (Kolen & Brennan, 1995). Cohen, Johnson, and Angeles (2000) considered the generation of common items as a "replicable process, including the selection of item writers and editors, the subdomains covered, the item types developed, and other such variables" (p. 6).

Responding to the Pellegrino et al. (1999) report, Cohen, Johnson, and Angeles (2000) framed the issue as a problem of sampling in two dimensions: "The assessment instrument contains a sample of items that reflect the construct or constructs of interest, and the study draws a sample of examinees who respond to these items" (p. 1). Standard errors reported for large-scale assessment programs quantify the latter but not the former source of variance, making the additional assumption that "items are virtually exchangeable when estimating aggregate statistics" (p. 2).

To our knowledge, the only study that addressed the issue of variability due to sampling of items was the unpublished study by Cohen et al. (2000). They applied a generalized version of the jackknife procedure (Quenouille, 1956) to evaluate the variance of estimates for the NAEP due to the two-dimensional sampling of examinees and items included in the assessment.[3] Cohen et al. (2000) found evidence of substantial underestimates of traditional NAEP standard errors. The uncertainty in IRT parameter estimates results in 25% to 100% increase in standard errors. The contribution of the sampling of examinees to the uncertainty in IRT parameter estimates was modest, whereas that of the sampling of items was more profound. Their results imply that as larger samples of examinees take an assessment, the standard error of measurement (and the variability in IRT item parameter estimates due to the sampling of examinees) will decrease, thus giving the impression of accurate measurement. The variability due to the sampling of items is unaffected by examinee sample sizes, and it could become the dominant source of error in large samples.

Error due to the selection of common items will affect both individual and aggregate scores, because equating functions rescale both onto a different scale. On an individual level, such error may not be as critical because there are large sources of uncertainty associated with individual scores. Aggregates, on the other hand, have much smaller standard errors, especially as samples of examinees become large. The standard error of equating that quantifies the uncertainty due to the sampling of examinees overestimates the accuracy of test scores by failing to consider the variability due to the sampling of common items. In this study, an analytical expression and a bootstrap approximation that quantify the latter source

[3] It should be noted that Cohen et al. (2000) considered the uncertainty in the IRT parameters due to the sampling of examinees and items in the complex NAEP design. NAEP's matrix-sampling uses a *balanced incomplete block design* and each examinee is administered only a few of the items in the pool.

of error are derived. The standard error due to the sampling of common items is compared to a small-sample bootstrap approximation of the standard error of equating due to examinee sampling to illustrate the relative magnitude of this source of error for individual and group-level interpretations.

Standard Error of Equating Due to the Selection of Common Items

Suppose that two groups of examinees, Groups 1 and 2, respond to two alternate forms of the same test under the common-item nonequivalent groups design. The two forms have $j = 1,..., p$ items in common. Each group's person-by-item response matrix is calibrated separately under a 3-Parameter Logistic (3PL) IRT model (Birnbaum, 1968).

$$P(x_{ij} = 1 \mid \theta_i, a_j, b_j, c_j) = c_j + \frac{1 - c_j}{1 + e^{-Da_j(\theta_i - b_j)}}$$

is the probability of examinee i to answer item j correctly, that is, to get a score of one, given his/her ability θ_i and the three item parameters a_j, b_j, and c_j. x_{ij} takes values of 0 and 1. D is a scaling constant equal to 1.7. Similar formulations can be given for 1- and 2-Parameter Logistic models.

The linear transformation used to place the score scale for Group 2 onto the scale of Group 1 is

$$\theta_{2i}^* = A\theta_{2i} + B$$

θ_{2i} are ability estimates for the $i = 1,..., N_2$ Group-2 examinees. θ_{2i}^* are the transformed Group 2 scores, equated to the Group 1 score scale.

Let $a_{2j}^* = \dfrac{a_{2j}}{A}$ and $b_{2j}^* = Ab_{2j} + B$ be the scaling transformations for the item parameters. Then it follows that for Group 2

$$P(x_{ij} = 1 \mid \theta_i, a_j, b_j, c_j) = P(x_{ij} = 1 \mid \theta_i^*, a_j^*, b_j^*, c_j)$$

which means that the linear transformation does not affect the probability of answering an item correctly or not. If the equating is accurate, however, values of θ_{2i}^* will be directly comparable (on the same scale) as values of θ_{1i}.

The mean/sigma method (Marco, 1977) is an IRT equating method that estimates the transformation constants A and B from the moments of the sample IRT difficulty values of the p common items for the two Groups as follows:

$$\hat{\mathrm{A}} = \sqrt{\frac{s^2(b_{1j})}{s^2(b_{2j})}} \text{ and}$$

$$\hat{\mathrm{B}} = \bar{b}_{1j} - \hat{\mathrm{A}}\bar{b}_{2j}$$

Because of model misspecification, and sampling of common items, there is variability in the estimates of A and B, which results in a source of error in the equated scores θ^*_{2i} due to the selection of common items[4]:

$$Var(\theta^*_{2i}) = Var(\hat{\mathrm{A}}\theta_{2i} + \hat{\mathrm{B}}) = Var\left[\sqrt{\frac{s^2(b_{1j})}{s^2(b_{2j})}}\theta_{2i} + \bar{b}_{1j} - \sqrt{\frac{s^2(b_{1j})}{s^2(b_{2j})}}\bar{b}_{2j}\right].$$

To evaluate this expression, the delta method is applied. Using the Taylor expansion around $\bar{b}_{1j}, \bar{b}_{2j}, s^2(b_{1j}), s^2(b_{2j})$,

$$Var(\theta^*_{2i}) = Var(\bar{b}_{1j})\left(\frac{d\theta^*_{2i}}{d\bar{b}_{1j}}\right)^2 + Var(\bar{b}_{2j})\left(\frac{d\theta^*_{2i}}{d\bar{b}_{2j}}\right)^2 +$$

$$Var\left[s^2(b_{1j})\right]\left(\frac{d\theta^*_{2i}}{ds^2(b_{1j})}\right)^2 + Var\left[s^2(b_{2j})\right]\left(\frac{d\theta^*_{2i}}{ds^2(b_{2j})}\right)^2 +$$

$$2Cov(\bar{b}_{1j}, \bar{b}_{2j})\frac{d\theta^*_{2i}}{d\bar{b}_{1j}}\frac{d\theta^*_{2i}}{d\bar{b}_{2j}} + 2Cov\left[s^2(b_{1j}), s^2(b_{2j})\right]\frac{d\theta^*_{2i}}{ds^2(b_{1j})}\frac{d\theta^*_{2i}}{ds^2(b_{2j})}.$$

Under the assumption of bivariate normality, which implies normality of the *b*′s for each group, the covariances of an average and a variance for sets of IRT *b*′s are zero and thus have been omitted from the above expression. To evaluate the variance of θ^*_{2i} two components are needed: the variances and covariances of $\bar{b}_{1j}, \bar{b}_{2j}, s^2(b_{1j}), s^2(b_{2j})$, and the derivatives of θ^*_{2i} with respect to each of $\bar{b}_{1j}, \bar{b}_{2j}, s^2(b_{1j}), s^2(b_{2j})$.

Let $\vec{s}_1$ be a q x 1 vector of parameter means and $\vec{s}_2$ the q(q+1)/2 x 1 vector of parameter covariance elements. In this case of two groups,

[4] The use of estimates of item parameters rather than the (unavailable) true parameter values also contributes error, but that source of error is commonly accounted for in examining the standard error of equating. The focus here is error due to sampling of common items.

$$\vec{s}_2^{\,t} = \left[\overline{b}_{1j}, \overline{b}_{2j}\right]$$

$$\vec{s}_2^{\,t} = \left[s^2\left(b_{1j}\right), s\left(b_{1j}, b_{2j}\right), s^2\left(b_{2j}\right)\right]$$

Under multivariate normality, a consistent estimator of the asymptotic covariance matrix of $\vec{s}^{\,t} = \left[\vec{s}_1^{\,t}, \vec{s}_2^{\,t}\right]$ is given by the symmetric matrix W

$$W \;= \begin{bmatrix} W_{11} & \\ W_{12} & W_{22} \end{bmatrix}$$

where the elements of the sub-matrices are

$$W_{11} = \frac{S}{N}$$

$$W_{12} = 0$$

$$W_{22} = \mathrm{K}(S\mathrm{x}S)\mathrm{K}' \text{ (Muthén, 1989).}$$

N is the sample size, S is the sample covariance matrix and K is a constant matrix selecting elements. The elements of W_{22} can be obtained by

$$(N-1)Cov(s_{ij}, s_{kl}) = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk} + \frac{N-1}{N}k_{ijkl}$$

where $k_{ijkl} = 0$ under normality.

In this case,

$$W_{11} = \frac{1}{p}\begin{bmatrix} s^2(b_{1j}) & \\ s(b_{1j}, b_{2j}) & s^2(b_{2j}) \end{bmatrix}$$

$$W_{22} = \frac{2}{p-1}\begin{bmatrix} s^4(b_{1j}) & & \\ s^2(b_{1j})s(b_{1j}, b_{2j}) & \dfrac{s^2(b_{1j})s^2(b_{2j}) + s^2(b_{1j}, b_{2j})}{2} & \\ s^2(b_{1j}, b_{2j}) & s^2(b_{2j})s(b_{1j}, b_{2j}) & s^4(b_{2j}) \end{bmatrix}$$

The derivatives of $\theta_{2i}^* = \sqrt{\dfrac{s^2(b_{1j})}{s^2(b_{2j})}}\,\theta_{2i} + \overline{b}_{1j} - \sqrt{\dfrac{s^2(b_{1j})}{s^2(b_{2j})}}\,\overline{b}_{2j}$ with respect to each of

$\overline{b}_{1j}, \overline{b}_{2j}, s^2(b_{1j}), s^2(b_{2j})$ are given by the following:

$$\frac{d\theta_{2i}^{*}}{d\bar{b}_{1j}} = 1$$

$$\frac{d\theta_{2i}^{*}}{d\bar{b}_{2j}} = -\sqrt{\frac{s^{2}(b_{1j})}{s^{2}(b_{2j})}}$$

$$\frac{d\theta_{2i}^{*}}{ds^{2}(b_{1j})} = \frac{\theta_{2i}}{2\sqrt{s^{2}(b_{1j})s^{2}(b_{2j})}} - \frac{\bar{b}_{2j}}{2\sqrt{s^{2}(b_{1j})s^{2}(b_{2j})}} = \frac{\theta_{2i} - \bar{b}_{2j}}{2s(b_{1j})s(b_{2j})}$$

$$\frac{d\theta_{2i}^{*}}{ds^{2}(b_{2j})} = -\frac{\theta_{2i}\sqrt{s^{2}(b_{1j})}}{2\left(\sqrt{s^{2}(b_{2j})}\right)^{3}} + \frac{\bar{b}_{2j}\sqrt{s^{2}(b_{1j})}}{2\left(\sqrt{s^{2}(b_{2j})}\right)^{3}} = \frac{(\bar{b}_{2j} - \theta_{2i})s(b_{1j})}{2s^{3}(b_{2j})}$$

Therefore, the formula to estimate the variance in the Group 2 equated scores is

$$Var(\theta_{2i}^{*}) = \frac{2s^{2}(b_{1j})}{p} + \frac{(\theta_{2i} - \bar{b}_{2j})^{2} s^{2}(b_{1j})}{(p-1)\, s^{2}(b_{2j})} - \frac{2s(b_{1j}, b_{2j})s(b_{1j})}{p\; s(b_{2j})} - \frac{(\theta_{2i} - \bar{b}_{2j})^{2}\left[s(b_{1j}, b_{2j})\right]^{2}}{(p-1)\, s^{4}(b_{2j})} \tag{1}$$

The square root of the expression in (1) gives an estimate of the standard error of equating due to the sampling of common items under the mean/sigma method for deriving the equating transformation.

Methodology

Data Sources and Calibrations

A mathematics assessment was administered to two statewide cohorts of Grade 8 students over two successive years: 2000-01 and 2001-02, referred to as Year-1 and Year-2 assessments respectively. The sizes of the two cohort populations were 7258 and 7128. There were a total of 139 and 137 items in the Year-1 and Year-2 administrations respectively, arranged in 8 forms. Forty-four items were embedded in both assessments, distributed across the 8 forms; they can be used as common items to link the two score scales.

Prior to equating Year-2 to Year-1, each assessment was calibrated separately with PARSCALE 3.0 (Muraki & Bock, 1997). The assessments came in multiple forms every year with certain items embedded in all those forms—in addition to the common items embedded across years for equating cohort scores. Concurrent calibration of all forms for a given year automatically places them on a single scale.

PARSCALE 3.0 allows calibration of tests that consist of both dichotomous and polytomous items and thus fitting of mixed models. First, a 3PL IRT model was fitted to the dichotomous data as in equation (2).

$$P_j(\theta_i) = c_j + (1 - c_j)\frac{e^{1.7a_j(\theta_i - b_j)}}{1 + e^{1.7a_j(\theta_i - b_j)}} \tag{2}$$

$P_j(\theta_i)$ is the probability that an examinee *i* with ability level θ_i responds to item *j* correctly. For each item *j*, the model provides estimates for the three parameters a_j, b_j, and c_j; the discrimination, difficulty or location, and pseudoguessing or lower asymptote parameters respectively. It also provides an estimate of the examinee's ability level θ_i. The dichotomous and the polytomous items were all scaled together. For the polytomous items, a graded response model (Samejima, 1969) was fitted. The graded response model applies to items that are scored in ordered categories with higher categories representing better performance than lower categories. The probability of an examinee scoring in a category *k* *(k=0,…,m)* is equal to the probability of obtaining a score of *k* or above, $P^+_{jk}(\theta)$, minus the probability of obtaining a score of *k+1* or above, $P^+_{j,k+1}(\theta)$. Its logistic form given by Muraki and Bock (1997) is

$$P_{jk}(\theta) = P^+_{jk}(\theta) - P^+_{j,k+1}(\theta) = \frac{e^{1.7a_j(\theta - b_j + d_k)}}{1 + e^{1.7a_j(\theta - b_j + d_k)}} - \frac{e^{1.7a_j(\theta - b_j + d_{k+1})}}{1 + e^{1.7a_j(\theta - b_j + d_{k+1})}} \tag{3}$$

where $P_{jk}(\theta)$ is the probability that an examinee with ability level θ obtains a score *k* on an item *j* with *m+1* scoring categories; d_k is the category parameter, with $\sum_{k=0}^{m} d_k = 0$, and $d_k - d_{k+1} \geq 0$. The difference $b_j - d_k$ is referred to as the *category-threshold* parameter, the location on the scale that separates two adjacent scores. The probability of responding in one of the two extreme categories or above is defined as $P^+_{j0}(\theta) = 1$ and $P^+_{j,m+1}(\theta) = 0$. A polytomous item has *m* category-threshold parameters (separating the *m+1* scoring categories), as opposed to a single location parameter for a dichotomous item. When a polytomous item appears in a common-item pool, it will contribute all its category-threshold parameters, and will thus have more weight in the equating transformation than the dichotomous items. This is not undesirable, since a polytomous item carries more weight (i.e., can give more score points) in the estimation of an individual examinee's score.

A 1PL model was also fitted,

$$P_j(\theta_i) = \frac{e^{1.7(\theta_i - b_j)}}{1 + e^{1.7(\theta_i - b_j)}}$$

by fixing the a_j parameters at 1 and excluding the c_j from the model. The graded response model (equation (3) with the a_j parameters fixed at 1) was fitted to the polytomous data.

Two estimates for the standard error of equating due to the selection of common items are presented under both a 3PL and a 1PL IRT model fit: (a) the analytical expression derived in the previous section, the square root of (1), and (b) a bootstrap approximation described in the following section. For the 1PL IRT case, this standard error is compared to the traditional standard error of equating due to imprecision in estimates of item parameters (i.e., due to examinee sampling), which is calculated using a small-sample bootstrap procedure.

Bootstrap Approximation to the Standard Error Due to the Sampling of Items

Data for the two annual administrations for the Mathematics 8 assessment were calibrated with both a 3PL and a 1PL IRT model, and item parameters were estimated from each model. The IRT *b* parameter estimates of the common items are relevant to the mean/sigma equating method. The slope A and the intercept B of the mean/sigma transformation are functions of the moments of the *b* estimates: $A = \frac{\sigma(b_{Jj})}{\sigma(b_{Ij})}$, and $B = \mu(b_{Jj}) - A\mu(b_{Ij})$ for groups I and J and common items j.

The particular assessment had a set of 44 common items: 37 dichotomous items and 7 polytomous items scored 0 to 4. In the original calibration, 65 estimates of item difficulty parameters entered the calculation of the mean/sigma transformation: 37 for the dichotomous and 28 for the polytomous (4 category-threshold estimates for each polytomous item).

The variability in the transformation was approximated with a bootstrap resampling technique (Efron, 1982; Efron & Tibshirani, 1993). The bootstrap procedure, which does not rest on any distributional assumptions, was applied as follows: Two thousand bootstrap samples were generated, each sample consisting of 44 common items drawn with replacement from the common-item pool. A mean/sigma equating transformation was estimated from each bootstrap sample using the Year-1 and Year-2 IRT difficulty values of the sampled common items; when a polytomous item was drawn, all four of its category-threshold estimates were used in the equating. The 2000 transformations resulted in 2000 corresponding scaled scores for each point θ of the theta scale. The standard deviation of 2000 scaled scores that correspond to a single θ is an approximation to the standard error of equating due to the selection of common items for that θ.

The magnitude of the standard error differs for each point on the scale. In the error variance formula (1), the terms for the difference between a point on the θ scale θ_{2i} and the mean of the IRT *b* parameter estimates are squared. Thus, the error curve has a parabolic shape indicating more accuracy at the center and less accuracy at the extremes of the scale. The variability in the equating transformation can be compared to the variability in an ordinary least squares regression line. The farther a point lies from the mean of the observations, the more uncertainty there is in the prediction of the dependent variable; the relative standing of the point is multiplied by the variability in the slope estimate. The variability in the intercept estimate is constant across the scale.

The analytic and the bootstrap methods estimate the same common-item sampling error. The former was derived assuming a bivariate normal distribution. As can be seen in the results section, the sets of IRT difficulty values showed substantial departures from bivariate normality. To establish whether departures from bivariate normality accounted for any discrepancy between the two methods, a simulation study was carried out as follows. Assume that the parent population from which data are sampled follows a bivariate normal distribution. The means, standard deviations, and covariance of this bivariate population are chosen to match the corresponding statistics of the original Mathematics 8 sample of IRT *b* estimates for Year 1 and Year 2. Two populations are assumed to exist: one with the aforementioned statistics equal to the sample statistics obtained under a 3PL calibration of the dichotomous items, and the second with statistics matching those of the sample in the 1PL case. Polytomous items are calibrated accordingly. Applying a parametric bootstrap technique (Efron & Tibshirani, 1993), 2000 samples of 65 pairs of observations are sampled from the parent population, as explained in the following paragraph. For each sample the mean/sigma equating transformation is calculated and each point on the θ scale is transformed to a different scale for equating purposes. The standard deviation of the 2000 equated scores for a theta is a bootstrap estimate of the common-item sampling error. The same procedure is repeated with the second parent population. These two parametric bootstrap estimates can be compared to the 3PL and 1PL analytic estimates, which are the same as in the original sample, since the same means, standard deviations, covariance and number-of-items terms are used in the variance formula (1), to assess whether they are more similar when the assumption of bivariate normality of the data holds.

The FORTRAN code for the parametric bootstrap simulation is presented in Appendix A. Sampling from a bivariate normal distribution with given moments was simulated as follows: A sum of 12 uniform variates minus 6 results in an approximately unit normal deviate, since the mean and variance of a (0,1) uniform distribution are $\frac{1}{2}$ and $\frac{1}{12}$ respectively. Sixty-five pairs $\left(X_{1i}, X_{2i}\right)$ of independent normal deviates were generated. The transformations $Y_{1i} = X_{1i}$ and $Y_{2i} = r\, X_{1i} + \sqrt{1-r^2}\; X_{2i}$ produce $\left(Y_{1i}, Y_{2i}\right)$ pairs that are correlated with a correlation coefficient *r*. After applying the linear transformations $Y_{1i}^* = \bar{x}_1 + s_1 Y_{1i}$ and $Y_{2i}^* = \bar{x}_2 + s_2 Y_{2i}$ the bivariate sample $\left(Y_{1i}^*, Y_{2i}^*\right)$ can be considered to have been drawn from a population with predetermined means $\bar{x}_1$ and $\bar{x}_2$, standard deviations s_1 and s_2, and correlation *r*.

Bootstrap Standard Error of Equating Due to the Sampling of Examinees

To compare the proposed standard error of equating with the standard error due to the sampling of examinees, another bootstrap approximation was employed. In the Year-1 data set for the Mathematics 8 assessment, there were 7258 examinee response strings.[5] A bootstrap sample of 7258 response strings was drawn from the original data set with replacement. The sample was calibrated using a 1PL IRT model and item IRT *b* estimates were extracted for the common items. The same procedure was followed for the Year-2 data set, which included 7128 response strings, with 7128 draws in this case. The two sets of common-item IRT *b*'s provided estimates for the slope A and intercept B of the mean/sigma transformation that rescaled the Year-2 scores to the Year-1 scale.

[5] A response string is a series of scored item responses for an individual examinee.

The process was repeated 50 times. It was a "small-sample" bootstrap approximation. Each sample involved a calibration through PARSCALE 3.0, a procedure that is time-consuming and occasionally presents convergence problems. However, 50 equatings based on 100 calibrations (50 for each year) are considered enough to provide a good approximation to the variability of the equating transformation. A 1PL IRT model was fitted to each bootstrap sample to minimize the number of calibrations that failed to converge. The 3PL IRT model does not converge as easily because it estimates more parameters than the 1PL. Even so, some bootstrap samples could not be calibrated. In the Year-1 original sample, only one examinee had attained a score in one category of a polytomous common item. Whenever the response string of that examinee was not drawn in a bootstrap sample, the calibration could not proceed. In those cases, new bootstrap samples were drawn that could be calibrated.

The 50 estimates for the slopes and intercepts, $\hat{A}_{boot}$ and $\hat{B}_{boot}$, where the subscript *boot=1,…, 50* stands for a replication of the parameter estimated from a bootstrap sample, can be found in Appendix B. The variation in equated scores is

$$Var(\theta_2^*) = Var(\hat{A}_{boot}\theta_2 + \hat{B}_{boot}) = \theta_2^2 Var(\hat{A}_{boot}) + Var(\hat{B}_{boot}) + 2\theta_2 Cov(\hat{A}_{boot}, \hat{B}_{boot}) \qquad (4)$$

The standard error of equating due to the sampling of examinees at each point θ_2 of the ability scale is the square root of equation (4). The variances and covariance of the slope and intercept estimates can be estimated by the 50 bootstrap replications of the equating.

Results

Figure 1(a) plots the standard error of equating that arises due to the selection of common items, when a 3PL IRT model was fitted to the dichotomous data. The standard error for each point on the theta scale on the horizontal axis is plotted in standard deviation units of the theta distribution. The distribution has a standard deviation of one. The two lines represent the analytic expression and the bootstrap approximation. The parabolic shape of the error graphs denotes that there is less uncertainty due to the selection of common items towards the middle of the ability scale than at the extremes. There is some discrepancy between the two methods, with the analytic approach showing less error at all points of the scale. The analytic approach rests on the assumption of bivariate normality of the IRT difficulty values, which is examined next.

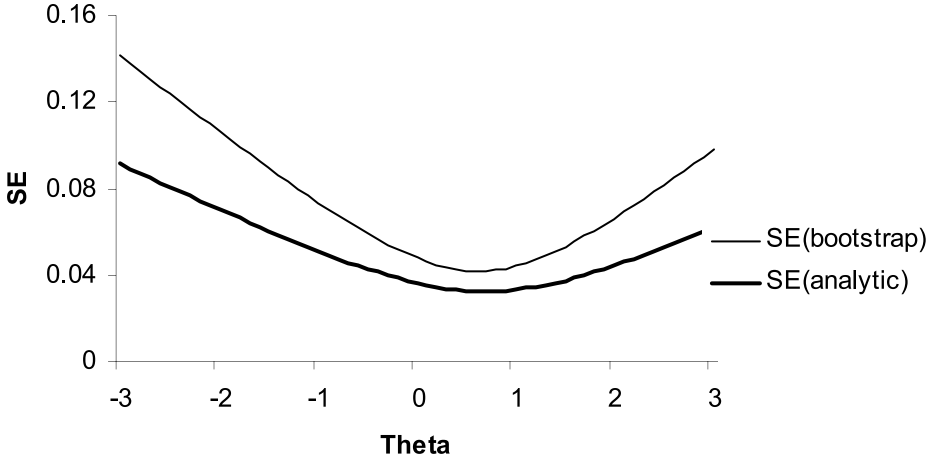

Figure 1(a). Standard error for common-item sampling for Mathematics 8 (3PL).

Figure 1(b) presents a similar graph when a 1PL IRT model was fitted to the data. At the lower part of the scale the analytic expression indicates more error than the bootstrap approximation, and less at the center and upper part.

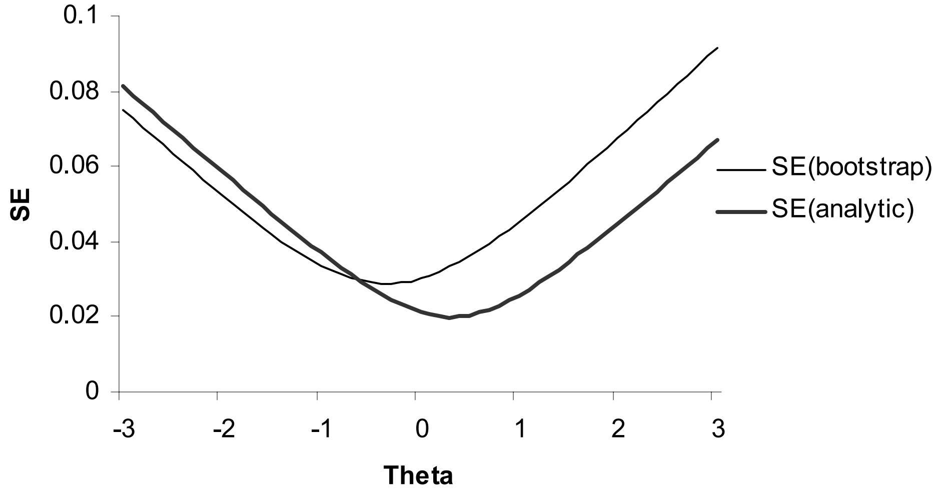

Figure 1(b). Standard error for common-item sampling for Mathematics 8 (1PL).

The variance formula (1) assumes a bivariate normal distribution of the sets of *b* parameter estimates used in the transformation. The bivariate plots of the IRT difficulty values appear in Figure 2. The difficulty values estimated with a 3PL calibration appear on the left panel and those from the 1PL calibration on the right panel. The Year-1 values are plotted on the horizontal axis and the Year-2 values on the vertical axis. There are 65 points on each plot representing 65 pairs of difficulty values entering the mean/sigma equating transformation. There is a high correlation of 0.98 between the Year-1 and Year-2 values under both models. However there is a large concentration of points around zero, which makes the univariate distributions of the *b* values leptokurtic. The univariate distributions of the *b*'s have a long right tail. In the 3PL plot there are two outliers in the upper end of the difficulty scale; those unusually high difficulties are the upper category-threshold estimates between the first, second, and third highest categories of a difficult polytomous item.

For a closer look at the normality assumption, Figure 3 presents the histograms for four sets of difficulty values: The Year-1 and Year-2 values obtained from a 3PL model appear on the left panel and those obtained from a 1PL model appear on the right panel. A normal curve is plotted in each case to enable comparisons. All four univariate histograms show departures from univariate normality. There are outlying values to the right tail of each distribution. Most of the difficulty values are concentrated at the middle of the distribution, resulting in a leptokurtic shape particularly for the 1PL case.

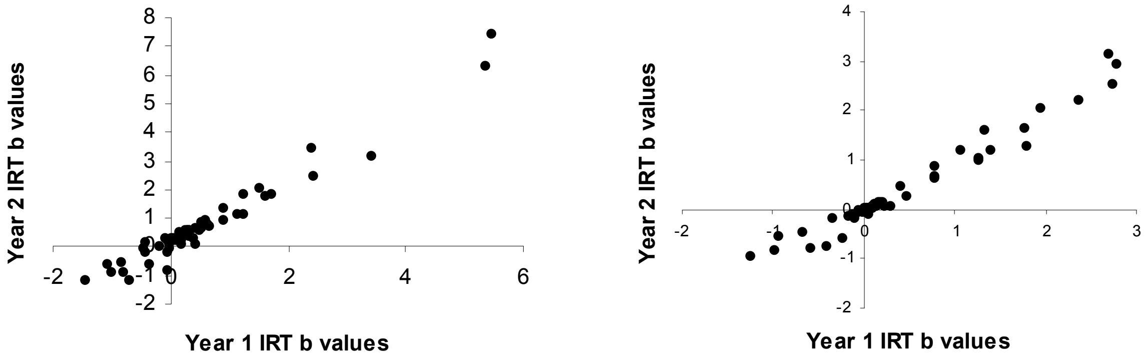

Figure 2. Bivariate plots of the common-item b values under a 3PL (left) and a 1PL (right) model.

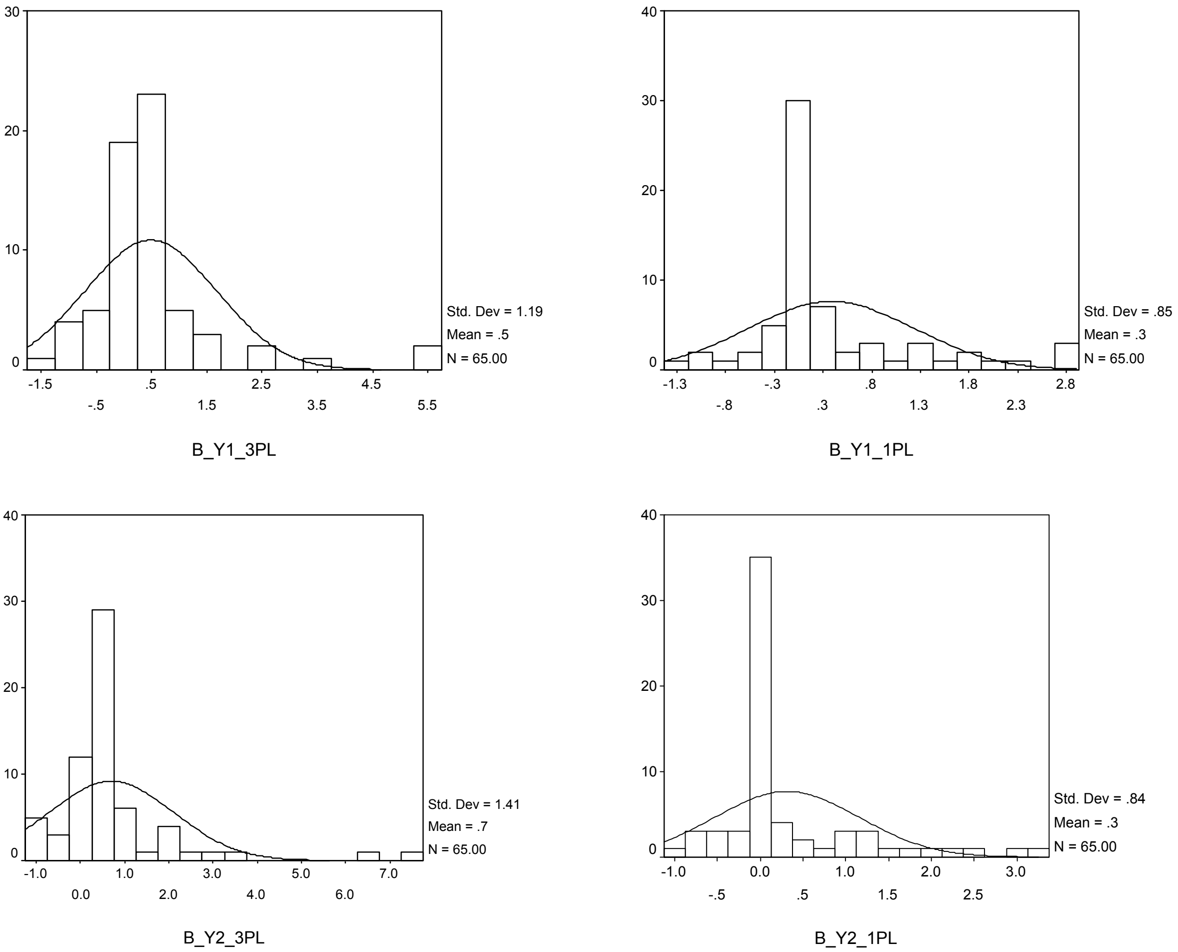

Figure 3. Histograms of IRT difficulty values by year and type of model fitted.

The Q-Q plots in Figure 4 confirm that the univariate distributions of the IRT *b* values depart considerably from normality. The quantiles of each observed set of *b* values do not match the quantiles of a normal distribution with the same mean and standard deviation as the observed distribution. All four plots have a characteristic shape with a steep slope at the center which demonstrates the peakedness of the distributions; the points below the diagonal line at the upper right side of each plot denote a heavy right tail, and the points below the diagonal at the lower left side of each plot denote a light left tail. The shapes of the distributions have a positive skew and a high kurtosis. Since the univariate plots for each set of *b* values show important departures from univariate normality, the assumption of bivariate normality, which underlies the analytic formulation of the standard error of equating due to the selection of common items, is violated as well.

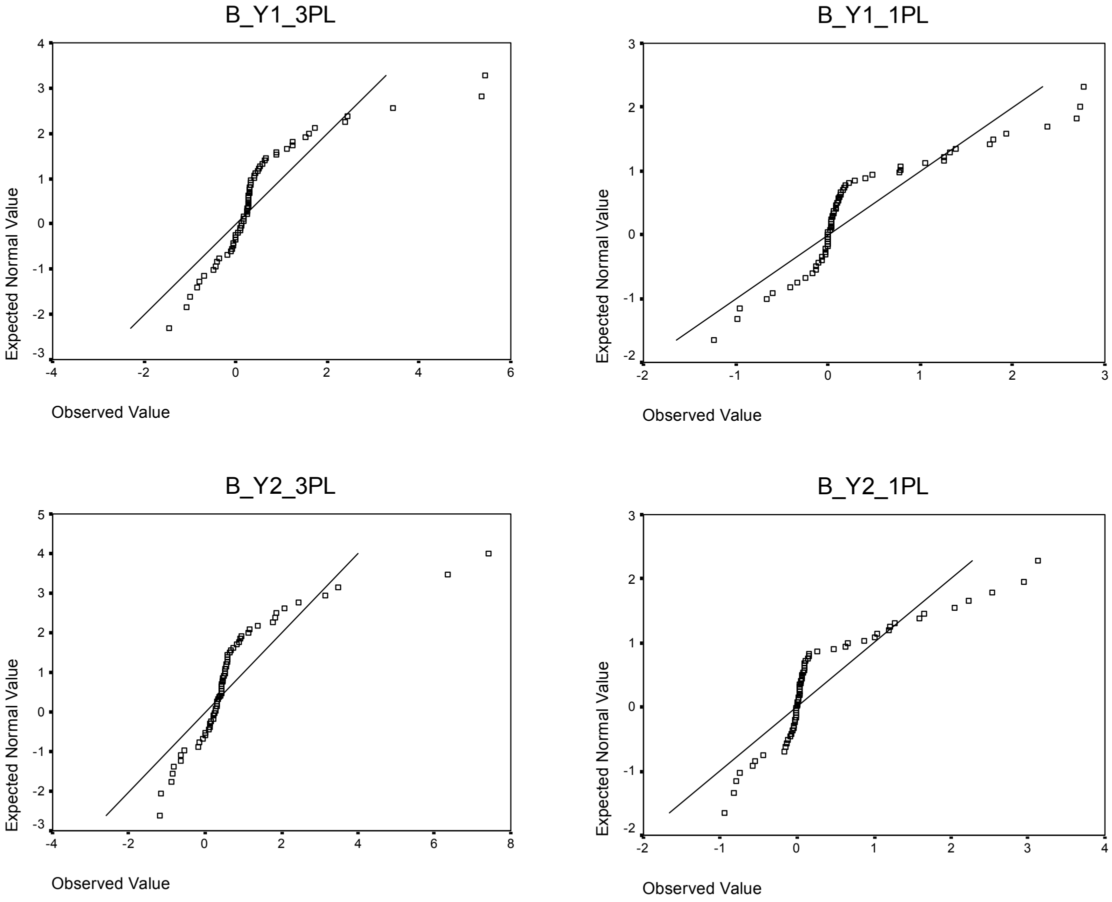

Figure 4. Q-Q plots for the IRT difficulty values by year and type of model fitted.

To determine whether the discrepancy in the estimation of the error between the two methods arises because of the departures from assumptions, the errors were computed for 65 pairs of observations that were drawn from a bivariate normal population. For this simulation, two parent populations, one for the 3PL case and one for the 1PL case, were assumed to have means, standard deviations and correlation equal to the corresponding values of the original data set. For the 3PL case the two sampled sets IRT *b* values had means 0.486811 and 0.697034, standard deviations 1.192396 and 1.409285, and a correlation coefficient of 0.975884. The corresponding values for the 1PL case were 0.335612, 0.311670, 0.847547, 0.840806, and 0.982069 respectively.

The results of the simulation appear in Figures 5(a) and 5(b). When the data are drawn from a bivariate normal population, the two methods give nearly identical

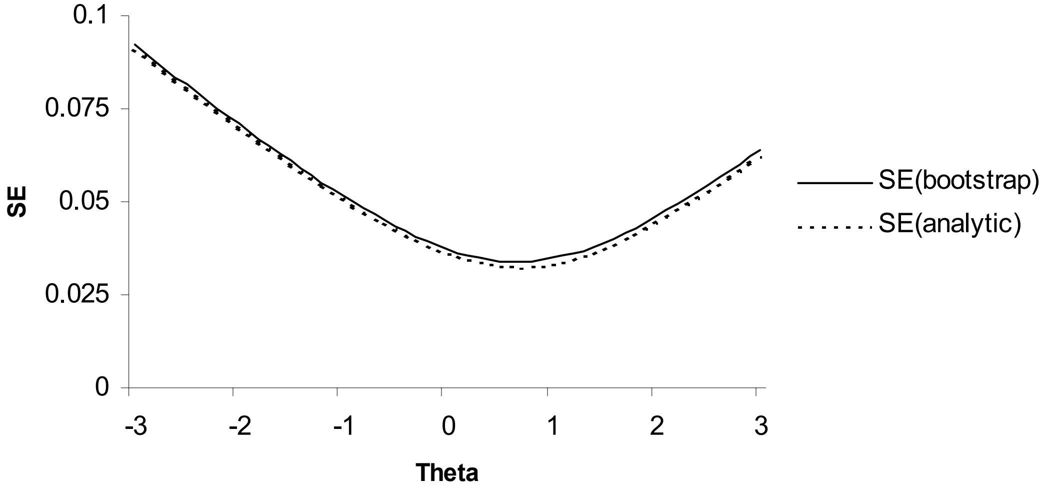

Figure 5(a). Simulated standard error for common-item sampling (3PL).

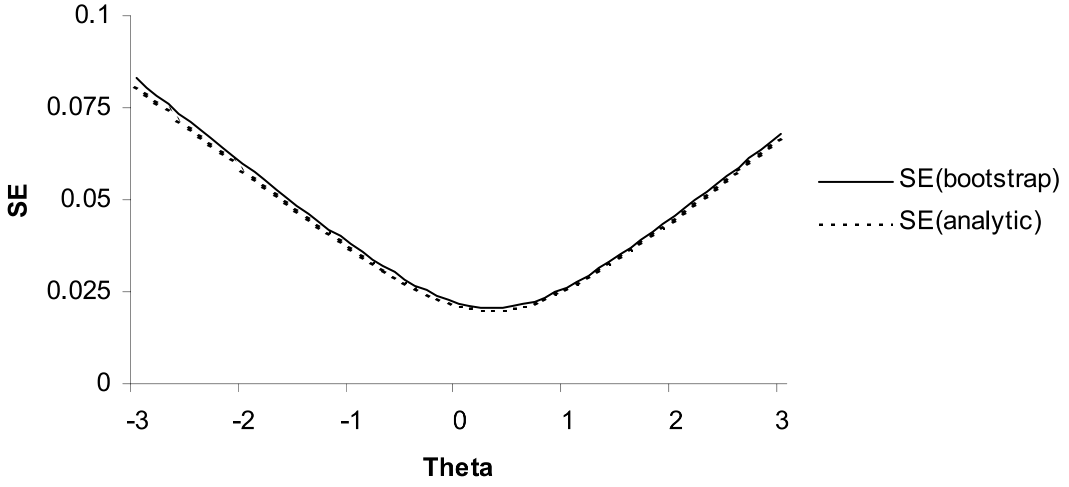

Figure 5(b). Simulated standard error for common-item sampling (1PL).

results. Departures from assumptions accounted for most of the discrepancy in the error graphs of the original data sets.

Figure 6 plots the two sources of standard error assuming a 1PL model. The analytic and bootstrap formulations of the standard error due to the sampling of common items can be compared to the traditional standard error of equating due to examinee-sampling variability. The two error sources are of similar magnitude. The graph of the latter is much more symmetric around the center of the theta scale, with

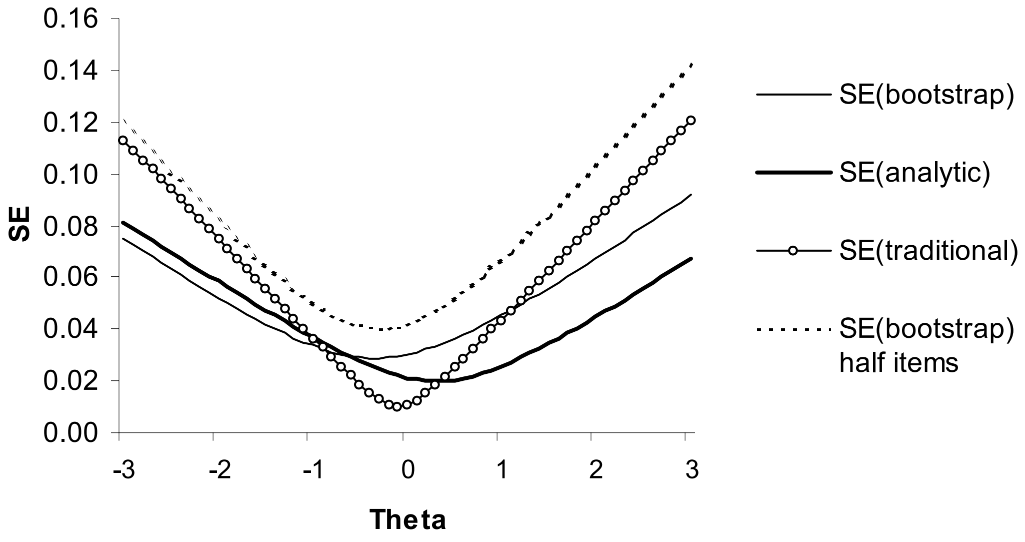

Figure 6. Comparison of error sources for the Mathematics 8 1PL calibration.

higher precision at the center and much less at the extremes. Figure 6 also presents a bootstrap error due to common-item sampling when only half of the common items (22) are sampled. With fewer common items, the variability due to common-item sampling is larger.

Discussion

Two ways of quantifying the variability in the equating transformation when the selection of common items is treated as a random process were demonstrated in this study. All standard error of equating graphs presented in this study had a parabolic shape showing that there is more uncertainty at the extreme scores of the ability distribution due to the variability of the equating transformation than at the center.

The closed formula derived for the variability due to common-item selection produced different results than the bootstrap approximation. Both methods gave standard errors that followed a similar pattern, but their magnitudes differed along the ability scale. Upon examination of univariate and bivariate normality of the item difficulty estimates it was found that there were violations to the assumption of bivariate normality, on which the analytic approach rests. In the simulation with data drawn from a bivariate normal population the two methods gave very similar estimates, implying that the analytic formula error estimates are distorted when the bivariate normality assumption does not hold. The bootstrap estimates are more defensible in the case of this particular assessment. Close agreement of the analytic results to the bootstrap results would appear to justify use of the simpler analytic approximation even in the presence of departures from bivariate normality of *b* parameter estimates; further work with more data sets should be undertaken before the formula can be recommended for general use.

The comparison of the proposed source of variance with the small-sample bootstrap approximation to the traditional source of equating error revealed that the two sources of error were of similar magnitude. This suggests that the currently ignored error variance could be important for the accuracy of equated scores reported. Its relative importance depends on the interpretation and use of scores (i.e., if they refer to individual examinee scores or group aggregates). Table 1 provides numerical examples of the size of different error sources for various points of the theta scale when the scores are used for individuals or for group means.

The three sources of error for individual scores are common-item sampling (under the bootstrap approximation), examinee sampling (which affects the equating transformation by way of its influence on item parameter estimates), and the standard deviation of the ability estimate $\hat{\theta}$. The values for the first two standard errors were obtained as described in the methodology section and are shown in Figure 6. The standard error for $\hat{\theta}$ is the estimate given for an examinee with ability $\hat{\theta}$ by PARSCALE 3.0, which is the standard deviation of the posterior distribution. Because in most cases there was no examinee with an integer ability estimate, the square root of the average squared standard deviations of the 10 neighboring $\hat{\theta}$ values serves as an approximate standard error of the reported $\hat{\theta}$. The rightmost column of Table 1 gives the sum of the squares of the separate error sources. Percentages are calculated in the metric of variances (squared errors). With regard to individual scores, then, the standard error of $\hat{\theta}$ is the dominant error source; at all five points of the scale it accounts for more than 90% of the total variability. Equating error, whether arising from the sampling of common items or the sampling of examinees, accounts for the remaining variance. Uncertainty for an individual score due to equating is very small compared to the uncertainty due to measurement imprecision especially at the center of the distribution.

Results are different when interpreting an aggregate score. The mean theta estimate for the Year-2 Mathematics 8 examinees was 0.014 with a 1PL calibration.

Table 1

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

	Source of standard error (Percentage of the total variance[a])			
$\hat{\theta}$	Common-item sampling	Examinee sampling	SE($\hat{\theta}$)	Total variance
For interpreting individual scores				
–2	0.05076 (2.2%)	0.07470 (4.9%)	0.32654 (92.9%)	
–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.

Table 1 tabulates the magnitude and relative size of different sources of error, for the case of a mean score of 0. The equating variability due to the sampling of common items and of examinees is the same as in the case of individual scores. The uncertainty in the accuracy of the mean due to sampling and measurement error—the standard error of the mean—is the square root of variance of the scores divided by the number of examinees[6]; it is very small with large samples of examinees, and estimates of mean scores tend to have high precision with large samples. Hence, the

[6] Although all students in the state in that particular grade took the test, there is still sampling error associated with the mean. The tested cohort can be conceived as a sample from an infinite population. For example, the population can be conceived as consisting of all the "potential past, current, and future examinees" (Kolen & Brennan, 1995, p. 213). Or that the outcome (e.g., annual mean gain) would appear with any examinee sample drawn from the same population, not just the particular sample (Cronbach et al., 1997).

equating errors, and in particular, the error due to common-item sampling, appear larger relative to the standard error of the mean. Common-item sampling error constitutes 82.6% of the total variance, a lot larger than the other sources of error, which are affected by the sample size. If a different point, away from the center of the distribution, were examined, the relative size of the two equating errors would shift. Error due to common-item sampling would be different for an estimate of the percent above a given cut score, for example. As shown in Figure 6, for the particular data set, at the extremes, the examinee sampling equating error would exceed the common-item sampling error. Since the assessment comes from a small state, for other, larger statewide testing programs the examinee sampling error could be even lower. This demonstration illustrates the magnitude of the additional variability introduced in aggregate scores in common-item sampling.

The sample sizes analyzed in this study were large enough, but in the context of statewide assessments, the samples came from a small state. Other statewide or national assessments are administered to larger samples of examinees. The standard error due to the sampling of examinees is inversely related to sample size; it becomes smaller as more examinee responses are used in the estimation, giving the impression of more precise measurement. The proposed source of error does not depend on sample size, so it could be much larger in those cases. It depends on the number of common items used for equating. In the Mathematics Grade 8 assessment there were 65 difficulty values representing 44 common items that entered the calculation of the equating transformation. Even with that many items, the uncertainty due to the sampling of items was comparable to the uncertainty in equating due to the sampling of the examinees.

This study employed the mean/sigma IRT equating method that uses item difficulty estimates. Variability in the equating transformations is pertinent to all methods under the common-item nonequivalent groups design, since the random versus fixed selection of common items is an issue for this particular equating design. Further research could be extended to develop ways to quantify the proposed standard error of equating when other equating methods are applied.

References

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Brennan, R. L., & Kolen, M. J. (1987). Some practical issues in equating. *Applied Psychological Measurement, 11,* 279-290.

Cohen, J., Johnson, E., & Angeles, J. (2000). *Variance estimation when sampling in two dimensions via the jackknife with application to the National Assessment of Educational Progress.* Washington DC: American Institutes for Research.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57,* 373-399.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap* (Monographs on Statistics and Applied Probability 57). New York: Chapman & Hall/CRC.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20,* 369-377.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6,* 195-240.

Holland P. W., & Rubin, D. B. (1982). Introduction: Research on test equating sponsored by Educational Testing Service, 1978-1980. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 1-6). New York: Academic Press.

Kolen, M. J. (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29-36.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1982). Standard error of an equating by item response theory. *Applied Psychological Measurement, 6*, 463-472.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14,* 139-160.

Muraki, E., & Bock, R. D. (1997). PARSCALE (Version 3.0) [Computer software]. Chicago, IL: Scientific Software International, Inc.

Muthén, B. (1989). Multiple-group structural modeling with non-normal continuous variables. *British Journal of Mathematical and Statistical Psychology, 42*, 55-62.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review (Otaru University of Commerce), 51*(1), 1-23.

Ogasawara, H. (2001a). Item response theory true score equatings and their standard errors. *Journal of Educational and Behavioral Statistics, 26*(1), 31-50.

Ogasawara, H. (2001b). Least squares estimation of item response theory linking coefficients. *Applied Psychological Measurement, 25,* 373-383.

Ogasawara, H. (2001c). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement, 25,* 53-67.

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (Eds.). (1999). *Grading the Nation's Report Card. Evaluating NAEP and transforming the assessment of educational progress.* Washington DC: National Academy Press.

Quenouille, M. H. (1956). Notes on bias estimation, *Biometrika, 43,* 353-360.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Iowa City, IA: Psychometric Society.

Tsai, T.-H. (1998). *A comparison of bootstrap standard errors of IRT equating methods for the common-item nonequivalent groups design*. Unpublished doctoral dissertation, University of Iowa, Iowa City.

Tsai, T.-H., Hanson, B. A., Kolen, M. J., & Forsyth, R. A. (2001). A comparison of bootstrap standard errors of IRT equating methods for the common-item nonequivalent groups design. *Applied Measurement in Education, 14,* 17-30.

Wainer, H. (1999). Comparing the incomparable: An essay on the importance of big assumptions and scant evidence. *Educational Measurement: Issues and Practice, 18*(4), 10-16.

Appendix A

FORTRAN Code for the Error Simulation
Assuming a Bivariate Normal Parent Population

```
PROGRAM ERROR_SIMULATION

! THIS PROGRAM CALCULATES BOOTSTRAP AND ANALYTIC STANDARD ERRORS DUE TO
! COMMON-ITEM SAMPLING ASSUMING THAT THE PAIRED OBSERVATIONS ARE DRAWN
! FROM BIVARIATE NORMAL POPULATION.
! CREATES A FILE WITH 3 COLUMNS: THETA POINT, BOOTSTRAP S.E., ANALYTIC S.E.

IMPLICIT NONE

REAL RHO,MEAN1,MEAN2,STDEV1,STDEV2,K,RAND,SET1(100),SET2(100),
REAL SQUARES1(100),SQUARES2(100),AVERAGE1,AVERAGE2,SD1,SD2,SLOPE(2000)
REAL INTERCEPT(2000),CONSTANT
REAL THETAS(61),SE(61),TRANSFORM(2000),TRANSQ(2000),ANALYTIC_VAR(61)
INTEGER N,TIMEARRAY(3),L,B

OPEN (UNIT=1,FILE='C:\ERROR\SIMULATE_BIVAR\ERROR8M3PL.TXT',STATUS='UNKNOWN')

! INITIALIZE VARIABLES TO ZERO
SQUARES1=0.0;SQUARES2=0.0;AVERAGE1=0.0;AVERAGE2=0.0;SD1=0.0;SD2=0.0
SLOPE(2000)=0.0;INTERCEPT(2000)=0.0
THETAS=0.0;SE=0.0;TRANSFORM=0.0;TRANSQ=0.0;ANALYTIC_VAR=0.0;CONSTANT=0.0

! ASSIGN VALUES FROM ORIGINAL DATA SETS TO VARIABLES
N=65                    !THE NUMBER OF NORMAL DEVIATE PAIRS
! FOR THE 3PL CASE
RHO=0.975884            !THE CORRELATION BETWEEN THE TWO SETS OF NORMAL DEVIATES
MEAN1=0.486811          !MEAN OF SET 1 OF NORMAL DEVIATES
MEAN2=0.697034          !MEAN OF SET 2 OF NORMAL DEVIATES
STDEV1=1.192396         !STAND. DEVIATION OF SET 1 OF NORMAL DEVIATES
STDEV2=1.409285         !STAND. DEVIATION OF SET 1 OF NORMAL DEVIATES

! FOR THE 1PL CASE
! RHO=0.982069
! MEAN1=0.335612
! MEAN2=0.31167
! STDEV1=0.847547
! STDEV2=0.840806

! GET CURRENT TIME TO INITIALIZE THE RANDOM NUMBER SEED
CALL ITIME(TIMEARRAY)
K=RAND(TIMEARRAY(1)+TIMEARRAY(2)+TIMEARRAY(3))

! BOOTSTRAP SAMPLING FROM THE PARENT POPULATION
DO B=1,2000
  SET1=0.0;SET2=0.0
  DO L=1,N

! CREATE TWO SETS OF APPROXIMATELY NORMAL DEVIATES
          SET1(L)=RAND(0)+RAND(0)+RAND(0)+RAND(0)+RAND(0)+RAND(0)+RAND(0)+RAND(0)+RAND
(0)+RAND(0)+RAND(0)+RAND(0)-6
          SET2(L)=RAND(0)+RAND(0)+RAND(0)+RAND(0)+RAND(0)+RAND(0)+RAND(0)+RAND(0)+RAND
(0)+RAND(0)+RAND(0)+RAND(0)-6
```

```
! TRANSFORM THE TWO SETS TO HAVE APPROPRIATE STATISTICS
        SET2(L)=RHO*SET1(L)+SQRT(1-RHO**2)*SET2(L)
        SET1(L)=MEAN1+STDEV1*SET1(L)
        SET2(L)=MEAN2+STDEV2*SET2(L)
  END DO

! DERIVE MEAN/SIGMA SLOPE AND INTERCEPT FOR EACH BOOTSTRAP SAMPLE
        SQUARES1 = SET1**2
        SQUARES2 = SET2**2
        AVERAGE1 = SUM(SET1)/N
        AVERAGE2 = SUM(SET2)/N
        SD1 = SQRT((SUM(SQUARES1)-SUM(SET1)**2/N)/(N-1))
        SD2 = SQRT((SUM(SQUARES2)-SUM(SET2)**2/N)/(N-1))
        SLOPE(B) = SD1/SD2
        INTERCEPT(B) = AVERAGE1-AVERAGE2*SD1/SD2
END DO

! CREATE 60 POINTS ON THE THETA SCALE FROM –3 TO +3
DO L=1,61
 THETAS(L)=-3.1 + 0.1*L
END DO

! FOR EACH THETA POINT GET THE BOOTSTRAP STANDARD ERROR [ SE(L) ]
! AND ANALYTIC STANDARD ERROR [ SQRT(ANALYTIC_VAR(L)) ]
DO L=1,61
 DO B=1,2000
   TRANSFORM(B)=THETAS(L)*SLOPE(B)+INTERCEPT(B)
 END DO
 TRANSQ=TRANSFORM**2
 SE(L)=SQRT((SUM(TRANSQ)-SUM(TRANSFORM)**2/2000)/(2000-1))

 CONSTANT=2*STDEV1**2/N-2*RHO*STDEV1**2/N
 ANALYTIC_VAR(L)=CONSTANT + (THETAS(L)-MEAN2)**2*(STDEV1**2/((N-1)*STDEV2**2)-
(RHO*STDEV1*STDEV2)**2/((N-1)*STDEV2**4))

 WRITE (1,'(T1,F8.5,T12,F8.5,T23,F8.5)') THETAS(L),SE(L),SQRT(ANALYTIC_VAR(L))
END DO

CLOSE(1)
END
```

Appendix B

Cycles to Convergence and Equating Constants for the Original and the Bootstrap Samples

Sample	Cycles to convergence in Year 1 data	Cycles to convergence in Year 2 data	Slope estimate	Intercept estimate
Original sample	8	23	1.0080	0.0214
Bootstrap samples				
1	11	23	0.9924	0.0120
2	8	26	1.0372	0.0374
3	8	23	1.0425	0.0372
4	8	23	0.9740	0.0115
5	8	26	1.0372	0.0374
6	8	23	0.9735	0.0232
7	8	26	1.0523	0.0303
8	8	23	1.0155	-0.0019
9	8	23	1.0387	0.0390
10	8	24	1.0445	0.0403
11	11	23	1.0100	0.0263
12	8	26	0.9966	0.0409
13	11	23	1.0409	0.0217
14	8	26	0.9841	0.0387
15	11	26	1.0644	0.0232
16	8	26	1.0372	0.0374
17	8	26	0.9841	0.0387
18	8	23	0.9834	0.0240
19	11	23	1.0154	0.0128
20	8	23	0.9830	0.0231
21	8	23	1.0387	0.0390
22	8	23	1.0194	0.0189
23	8	23	0.9740	0.0115
24	8	23	0.9463	0.0276
25	10	26	0.9908	0.0338
26	11	23	1.0456	0.0034
27	8	26	1.0994	0.0331
28	8	26	0.9961	0.0152
29	8	23	0.9759	0.0157
30	11	23	1.0348	0.0104
31	14	23	1.0192	0.0243
32	8	26	1.1036	0.0317
33	8	23	0.9735	0.0232
34	8	26	1.0352	0.0365
35	14	23	1.0192	0.0243

Sample	Cycles to convergence in Year 1 data	Cycles to convergence in Year 2 data	Slope estimate	Intercept estimate
36	8	26	1.0372	0.0374
37	11	26	0.9876	0.0207
38	11	24	1.0272	0.0220
39	8	23	0.9740	0.0115
40	11	23	0.9638	0.0165
41	11	26	0.9311	0.0209
42	8	23	1.0112	0.0211
43	8	26	1.0523	0.0303
44	11	23	1.0409	0.0217
45	11	26	0.9552	0.0190
46	11	26	1.0644	0.0232
47	8	23	0.9740	0.0115
48	8	26	1.0406	0.0452
49	11	23	1.0223	0.0132
50	11	26	0.9336	0.0200