A Study of Equating in NAEP

Larry V. Hedges
University of Chicago

Jack L. Vevea
University of North Carolina

Commissioned by the NAEP Validity Studies (NVS) Panel
December 1997

George W. Bohrnstedt, Panel Chair
Frances B. Stancavage, Project Director

The NAEP Validity Studies Panel was formed by the American Institutes for Research under contract with the National Center for Education Statistics. Points of view or opinions expressed in this paper do not necessarily represent the official positions of the U. S. Department of Education or the American Institutes for Research.

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

For Information:

NAEP Validity Studies (NVS)
American Institutes for Research
1791 Arastradero Road
PO Box 1113
Palo Alto, CA 94302
Phone: 415/493-3550
Fax: 415/858-0958

Acknowledgments

This study was conducted under the auspices of the NAEP Validity Studies (NVS) Panel. We thank the members of the NVS Panel, particularly Albert Beaton, Darrell Bock, George Bohrnstedt, and Fran Stancavage for helpful comments on previous drafts of this paper.

Contents

List of Tables

List of Figures

Abstract

This study investigates the amount of uncertainty added to NAEP estimates by equating error under both ideal and less than ideal circumstances. For example, circumstances led to a situation in which the 1994 to 1992 reading assessment equating had to be based on a set of common items that was both smaller, and more heavily weighted toward multiple choice, than anticipated. If performance on the two types of items does not change at the same rate over time, such equatings might introduce systematic bias in trends measured from equated scores. Data from past administrations are used to guide simulations of various (better and worse) equating designs, and error due to equating is estimated empirically.

The design includes a variety of factors that might affect accuracy of equating, with the levels of each factor based roughly on operational values in the NAEP 1992 and 1994 reading and 1992 mathematics assessments. The purpose is to estimate the approximate additional uncertainty that might be introduced by equating from one assessment wave to the next, and to determine what factors in the equating design contribute most to that uncertainty. The specific factors investigated were number of items in the scale, the proportion of items in the scale taken by each student, the proportion of items in each administration which are common, the proportion of each item "type" in each scale, the proportion of each item type among common items used for equating, the scale linking strategy (IRT invariance, common item, or multiple group IRT linking), and the change in ability from wave 1 to wave 2.

Common item scale linking performed very well, even under circumstances which were far from ideal, including slight to moderate multidimensionality. Mean bias was estimated to be no more than about 0.01 to 0.02 standard deviations (about 0.5 to 1.0 NAEP scale points). However, in nonideal conditions there were biases in the extreme quantiles (5 percent, 10 percent, and 25 percent points) of the ability distribution, even with no population shifts. These biases were several times as large as the mean bias and could be large enough to create problems in tracking low performance and the means of low performing groups over several waves of assessment. When both waves of data can be scaled together, multiple group IRT methods provided very accurate scale linking, with virtually no bias.

Introduction

In this study we examined the problem of equating error in NAEP-like assessment designs with complex samples and conditioning with multiple imputation, under conditions that closely resemble those of an operational assessment. The simulations were based on characteristics of the 1992 and 1994 NAEP reading and mathematics assessments.

A major question was how to incorporate "real data"—that is, characteristics of the actual assessment—into the simulation design. Several approaches were considered. One approach is to use item response strings from real respondents. This has the advantage of producing absolutely real data. It would incorporate, for example, the degree to which real item responses fail to conform to the item response model used in the analysis. However it has the important disadvantage that the generating parameters (other than scale length, number of items taken and number of items that are "common") are not under the control of the investigator and cannot be exactly known.

Another alternative is to use real data to derive reasonable values of person ability and item parameters and then simulate item response strings based on those (known) parameter values. This has the advantage of complete control over all relevant parameters (and knowledge of their values). If the values of these parameters are derived from estimates in operational assessments, they should be a good approximation to reality. However, real data may fail to conform to our analytic models in ways we do not fully understand (e.g., they may not fit the item response or multiple imputation model). This procedure has the disadvantage that it cannot capture the consequences of the misfit of real data to our analytic models. One might see the latter approach as suggesting a lower bound for errors when the rest of the model fits *exactly*.

We decided to use this latter approach—to simulate data that fit the item response model rather than use item response strings from real people. Specifically, we used the distribution of person ability parameters and caseweights obtained in the 1992 and 1994 NAEP reading and mathematics assessments, item parameters selected from the values for items in these same NAEP assessments, and the correlation between background variables (used in conditioning) and ability scores observed in the 1992 NAEP assessment.

Population. Two populations of person ability parameters were used in this study: one derived from the 1992 NAEP mathematics assessment and the other from the 1994 NAEP reading assessment. They were derived by taking the abilities from a random sample of 4,000 cases each from the 1992 NAEP mathematics and the 1994 NAEP reading samples for 17–year–olds. The average of each person's five plausible values served as the generating values of the person ability populations. The weights for these cases were preserved for the analysis as well.

Items. For the reading study, we used a sample of the item parameters from the 1992 reading assessment for 17–year–olds to serve as the generating item parameters. For the

mathematics study, we used the item parameters of the 1992 NAEP mathematics assessment for 17–year–olds. All item parameters were taken from the 1992 NAEP technical manual.

Design

The rationale for the design is guided by features of operational NAEP and the implementation of the short–term trend studies as part of the main assessment. Seven factors defined the conditions we investigated initially:

Total number of items. Although the overall item pool in NAEP is large, scaling is carried out within individual scales that have relatively small numbers of items. For example, the 1994 NAEP reading scales ranged from 20 to 40 items and the 1992 NAEP mathematics scales ranged from 21 to 47 items. Two scale lengths are examined in this simulation—a short scale of 24 items and a long scale of 48 items.

Proportion of items taken by each student. In order to obtain information about a range of items, each student who takes items on a particular scale takes only a fraction of the total NAEP item pool on each scale. For example, in the 1994 reading assessment, a student typically took one or two reading blocks corresponding to about one fourth of the items on a scale (if only one item block corresponding to a scale was taken) or one half of the items on a scale (if two blocks corresponding to the same scale were taken). This simulation examined two situations corresponding to every student taking one fourth or one half of the total number of items on the scale.

Proportion of items treated as common in equating. Although the same items are used for each wave of the short–term trend studies in NAEP, not all of these items are treated as common for the purposes of equating one assessment wave to the next. When the parameters of an item drift too much from one assessment wave to the next, those items are not included among the "common" items used for equating. Such drifting of item parameters is more likely to occur for constructed response items where there has been a change in the scoring procedures. For example, in the 1992 to 1994 reading short–term trend analysis within the main assessment, improvements in the scoring procedures for constructed response items led to decisions that only 57 percent to 85 percent of the items could be used as common items for the purposes of equating. Like NAEP, the simulation reported here used the same items for both waves of the assessment, but examined two situations: one where 50 percent of the items were treated as common in the equating and the other where 100 percent of the items were treated as common in the equating.

Proportion of Type I items. There are two types of items in NAEP—multiple choice items and constructed response items. Constructed response items are further subdivided into short constructed response items (which are scored dichotomously, but with the

guessing parameter set to zero) and extended constructed response items (which are scored using a partial credit model). Overall, constructed response items made up from 47 percent to 81 percent of the total items in the three reading scales used in the 1994 NAEP reading assessment. In this simulation two types of items are included, which are labeled Type I and Type II, corresponding to multiple choice and short (dichotomously scored) constructed response items, respectively. Two scale types were investigated, one with relatively few (50 percent) Type I items and with a larger proportion (two–thirds to three–quarters) of Type I items. The exact proportions were varied somewhat to accommodate other factors in the design.

The notion that Type I and Type II items measured slightly different ability dimensions was realized by using the model underlying the Bock, Gibbons, and Muraki (1988) full information item factor analysis model. Let θ_{1i} be the ability of the *i*th person on the first ability dimension (corresponding to what is measured in common by Type I and Type II items) and let θ_{2i} be the ability measured only by Type II items. If Type I and Type II items correspond to multiple choice and constructed response items, respectively, then θ_{1i} might correspond to a dimension of general knowledge and θ_{2i} might correspond to a production dimension measured only by constructed response items. The operational ability for person *i* is $\lambda_j\theta_{1i} + (1-\lambda_j)\theta_{2i}$, where the value of λ_j is determined by the type of item. In this simulation we used the value $\lambda = 1$ for Type I items and $\lambda = .9$ for Type II items.

Only dichotomously scored (short) constructed response items were examined in this simulation for two reasons. The first is conceptual. The vast majority of constructed response items are of the short constructed response type. For example, 80 percent of the constructed response items in the 1994 NAEP reading assessment for 17–year–olds and 87 percent of the constructed response items in the 1992 NAEP mathematics assessment for 17–year–olds were short constructed response items. Moreover, the extended constructed response items actually showed less of a tendency to drift than did the short constructed response items in the 1992 to 1994 reading short–term trend analysis. The second reason that extended constructed response items were not used in the simulation was that it would have required software that was not available to us (a NAEP proprietary program combining Bilog and Parscale).

Proportion of Type I items treated as common for equating. In the 1992 to 1994 NAEP reading short–term trend analysis within the main assessment, the items that were not included as common items used for equating were exclusively constructed response items. Consequently, although the scales were composed of between 47 percent and 71 percent constructed response items, the common items used for equating had a much smaller proportion of constructed response items, between 13 percent and 63 percent. It is unclear what effect on equating might arise when the items on a scale are predominantly of one type (e.g., constructed response) but the items treated as common for the purposes of equating are predominantly of another type (e.g., multiple choice). In this simulation, the proportion of Type I items used as common items for equating ranged from 16.7 percent to 50 percent.

Type of equating and scale linking. Two alternative strategies for equating assessment waves and linking scales were investigated in the main simulation: one based on strict IRT invariance (which has been proposed for, but is not used in, operational NAEP) and the other based on common item linking, which is similar to the strategy used in operational NAEP. In addition, we investigated a new strategy for equating and linking based on multiple group IRT (Bock and Zimowski, 1997).

Change in ability from one assessment wave to the next. One of the problems that contributes to the difficulties in linking scales in NAEP is that the ability distribution is changing from one assessment wave to the next. For example the change from 1992 to 1994 in reading for 17–year–olds was about 0.12 standard deviations. In this simulation we examined changes of 0.0 and 0.15 standard deviations between assessment waves. In initial trials of this simulation, both ability dimensions (θ_1 and θ_2) were changed the same amount. However, because change alone was not the primary interest, but *differential* change in the two ability dimensions, changes were introduced in the simulation so that the change in ability was only 90 percent as great on Type II items as on Type I items. Thus the change introduced was -0.15 on Type I items, but it was only -0.135 on Type II items.

The seven, two–level factors in the design description yield a total of 128 combinations in the completely crossed design. Initial investigation suggested that 80 of these cells were of most interest in that they posed the substantial challenges to equating and scale linking. Consequently our analyses and reporting have concentrated on these 80 combinations of factors. These 80 combinations can be most easily described in terms of 20 cells defined by the first five design factors, crossed with the final two factors. We will refer to the cells defined by the first five factors in the results that follow.

The following table describes the item layout that was generated for each replication of each cell of the design defined by the first five factors described above.

Item Layout: First Five Design Factors

	Factor				
	1 Scale Total Items	2 Items Taken	3 Total Common Items	4 Type I Items	5 Common Type I Items
Cell					
1	48	12	6	6	1
2	48	12	6	6	3
3	48	12	6	7	1
4	48	12	6	8	3
5	48	12	12	6	6
6	48	24	12	12	2
7	48	24	12	12	6
8	48	24	12	14	2
9	48	24	12	16	6
10	48	24	24	12	12
11	24	6	3	3	1
12	24	6	3	3	2
13	24	6	3	4	1
14	24	6	3	4	2
15	24	6	6	3	3
16	24	12	6	6	2
17	24	12	6	6	4
18	24	12	6	8	2
19	24	12	6	8	4
20	24	12	12	6	6

Scale Linking

In every scale linking there is a calibration step and a *scaling* step. The calibration step involves obtaining item parameter estimates from a computer program. As in NAEP, Bilog was used to do this calibration. The scaling step takes the output from the calibration program and turns it into scaled proficiency scores. This step is always a linear transformation, so if T_i is the scaled proficiency score and θ_i is the output from the computer program,

$$T_i = A + B\theta_i.$$

There are two waves of test administrations: 1 and 2.

There are also potentially three sets of item parameter estimates: those based on calibrating data from administration 1, those based on calibrating data from administration 2, and those based on calibration of the common data across the two administrations. Call these sets of item parameters 1, 2, and C respectively.

Denote the ability score for person *i* in administration (time) j, estimated from parameter k by $\theta_i(j, k)$, where the *i*th person in assessment wave 1 is not the same individual as the *i*th person in wave 2. Then we have:

> $\theta_i(1,1)$: the ability of the *i*th person wave 1, estimated using item parameters from the wave 1 calibration
>
> $\theta_i(1,C)$: the ability of the *i*th person in wave 1, estimated using item parameters from the common calibration
>
> $\theta_i(2,C)$: the ability of the *i*th person in wave 2, estimated using item parameters from the common calibration.

Define the moments of the $\theta_i(j, k)$ via:

> $M(j, k) = E_W[\theta_i(j, k)]$ (the weighted sample mean)
>
> $S(j, k) = \sqrt{VAR_W[\theta_i(j, k)]}$ (the weighted standard deviation).

Denote the *scaled* scores corresponding to the above ability estimates as for θ_i above except with a capital T instead of θ, e.g., $T_i(j, k)$ instead of $\theta_i(j, k)$.

Note that we need to define the scaling parameters A and B to define the linking. We start with a scale defined by A and B which are given *a priori*. Note that A and B are the mean and standard deviation of the scale if $M(1,1)=0$ and $S(1,1)=1$, but there may be cases when this is not true (as when wave 1 data has been previously linked to an earlier scale).

IRT Invariance Linking

In IRT invariance linking we link scale 1 to scale 2 as follows:

> $T_i(2,1) = A + B*\theta_i(2,1)$.

Note that this notation implies that all common items in wave 2 are constrained to have their wave 1 parameters and the parameters of noncommon items are unconstrained. This

is perhaps the strictest version of IRT invariance linking in that item drift (among items taken as common for the purposes of equating) is uncontrolled and is therefore confounded with change in ability. Other linking possibilities exist. For example, the method of Stocking and Lord (1983) imposes an item drift model by constraining the mean and variance of the item parameters, but allowing the particular values of item parameters to drift subject to this overall constraint. We evaluated this strict IRT invariance model because it represents one extreme which performed surprisingly well in the study of NAEP equating performed by Mazzeo and Donoghue (1995).

Common Item Linking

In common item linking we link scale 1 to scale 2 as follows:

$T_i(2,C) = A_{CI} + B_{CI}*\theta_i(2,C)$, where we derive A_{CI} and B_{CI} by making sure that the mean and variance of $T_i(1,C)$ are equal to those of $T_i(1,1)$.

Note however that there is no reason to believe that M(1,C)=0, even if M(1,1)=0, since the common calibration will change item parameters and therefore the ability scores. Similarly, there is no reason to believe that S(1,C)=1 even if S(1,1)=1.

Since $T_i(1,1) = A + B\theta_i(1,1)$ has mean A + B*M(1,1) and standard deviation B*S(1,1), then it follows that $\{[\theta_i(1,C) - M(1,C)]/S(1,C)\}*B*S(1,1) + A + B*M(1,1)$ also has mean A + B*M(1,1) and standard deviation B*S(1,1), since the term in brackets is just a z–score. Collecting terms we get that:

$A_{CI} = A + B*M(1,1) - [M(1,C)/S(1,C)]$, $B_{CI} = B*S(1,1)/S(1,C)$.

Multiple Group IRT Linking

Multiple group IRT provides an alternative to the two linking strategies outlined above (Bock and Zimowski, 1997). Multiple group IRT makes it possible to simultaneously scale the items in several populations, using the distribution of ability in one of the populations to anchor those of the other populations and providing automatic scale linking. When several populations (e. g., several waves of trend data) can be scaled at once, this method has theoretical advantages over the other models considered here.

Item Types

We will assume that the items actually follow a variant of the multidimensional item response model given in Bock, Gibbons, and Muraki (1988), where there are two ability dimensions and a three parameter logistic item response model. In this model, the items function as if there was a single ability which is a linear combination of the two individual ability dimensions. The coefficients of this linear combination determine the factor loading of each item on the ability factors. Thus if θ_{1i} and θ_{2i} are ability scores on the two abilities for person i, person *i* can be treated as if there were a single ability factor and the ability score for item *j* was:

$$\theta_{\cdot} = \lambda_{1j}\theta_{1i} + (1-\lambda_{1j})\theta_{2i}.$$

Thus to simulate Type I and Type II items, we generate two independent abilities such that the $\theta_{\cdot}$ values have the required distribution while letting λ_{1j} take on one value for Type I items and another value for Type II items. We generate the θ_{1i} and θ_{2i} values by assuming that they are uncorrelated, but that both abilities are equally correlated with the background variables.

Multiple Imputation

In our simulation study we are able to make certain simplifying assumptions that make computations easier. Since we estimate only one scale at a time, our ability scores are univariate, not multivariate as in the main NAEP. Similarly, we can treat the background variables as a single variable (one optimal composite of all the background variables). This section is an attempt to clarify the procedures and the notation we will use:

x_i—the *i*th person's item response string, with elements $x_i' = (x_{ij})$,

y_i—the *i*th person's background characteristics (all rolled into one variable),

θ_i—the *i*th person's ability parameter,

γ—the slope coefficient linking θ_i and y_i,

σ—the residual standard deviation in the above regression.

The posterior distribution of θ_i given $\mathbf{x}_i$, y_i, γ, and σ is given by:

$$p(\theta_i \mid \mathbf{x}_i, y_i, \gamma, \sigma) = P(\mathbf{x}_i \mid \theta_i, y_i, \gamma, \sigma)p(\theta_i \mid y_i, \gamma, \sigma).$$

Since the item response model says $\mathbf{x}_i$ depends only on θ_i, it follows that:

$P(\mathbf{x}_i \mid \theta_i, y_i, \gamma, \sigma) = P(\mathbf{x}_i \mid \theta_i)$ = Product over items $p(x_{ij} \mid \theta_i)$, where $p(x_{ij} \mid \theta_i)$ is just the probability that person *i* gets item *j* correct, which is given as a function of θ_i by the logistic IRT model.

The conditioning model says that θ_i depends on background variable y_i via a linear regression. We can standardize y_i, and θ_i is already standardized, so:

$\theta_i = \gamma y_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$.

Therefore $p(\theta_i \mid y_i, \gamma, \sigma) = \phi((\theta_i - \gamma y_i)/\sigma)/\sigma$, where ϕ is the standard normal probability density function. In the univariate case γ = correlation(y_i, θ_i), and $\sigma^2 = 1 - \gamma^2$, so σ determines γ.

We use NAEP's multiple imputation process (which includes the conditioning on background variables). The process has three steps:

1. Draw a value of γ from the normal approximation of $p(\gamma, \sigma \mid \mathbf{x}_i, y_i)$ fixing σ at its mean. We set σ, so we know its value and since σ determines γ entirely, this step is trivial.

2. Given γ and σ and y_i, get the maximum likelihood estimate of the mean and variance μ_p and $\sigma_p 2$ of the posterior distribution of θ_i given $\mathbf{x}_i$, y_i, γ and σ.

3. Sample 5 θ_i values from a normal distribution with mean μ_p and variance $\sigma_p 2$—these are the plausible values.

Generating Values for Multiple Imputation

The NAEP technical manual reports the amount of variance in the θ_i's that the background variables account for (the R^2 values) in the 1992 NAEP analysis. In reading, the proportion of variance accounted for is 0.40 (based on 39 conditioning variables) in the long–term trend and about 0.58 for each of the three reading scales (based on 115 principal components from 218 background variables) in the main assessment. In mathematics the proportion of variance accounted for by the background variables ranged from 0.20 to 0.31 for the five mathematics scales in the main assessment (based on 138 principal components from 238 background variables). Thus an R^2 of 0.25 (for math) and 0.52 (for reading) were chosen for the (squared) correlation between background variables and θ_i.

Simulation Methods

Overview

To the degree possible, control of the simulations was automated. We directed the sequence of program runs needed to complete the study of test equating by automatically generating batch files that called the necessary executables for simulation of data, calibration, equating, generation of plausible values, and assessment of the plausible values' distribution. We used the public release of Bilog 3 (as described in Mislevy & Bock, 1990) for all scaling except that in the multiple group IRT scaling analyses, where we used Bilog–MG (as described by Bock and Zimowski, 1997). The other steps employed programs written specifically for this project in the C programming language. Within a particular cell of the design, we performed the following steps to evaluate IRT invariance equating and common item equating:

1. Generate data for original (wave 1) assessment.
2. Calibrate the data, using Bilog.
3. Generate plausible values for original (wave 1) assessment.
4. Assess the distribution of the plausible values.
5. Generate wave 2 data, assuming no change in ability.
6. Calibrate wave 2 data, using IRT invariance equating strategy.
7. Generate plausible values.
8. Assess the distribution of the plausible values.
9. Calibrate wave 2 data, using common item equating strategy.
10. Generate plausible values.
11. Assess the distribution of the plausible values.
12. Generate new wave 2 data, assuming a change in ability (θ_1) of -0.15 standard deviations.
13. Calibrate the new wave 2 data, using IRT invariance equating strategy.
14. Generate plausible values.
15. Assess the distribution of the plausible values.

16. Calibrate the new wave 2 data using common item equating strategy.

17. Generate plausible values.

18. Assess the distribution of the plausible values.

For multiple group IRT equating we generated new wave 1 and wave 2 data, equated using Bilog-MG, and generated and analyzed plausible values as before.

In the paragraphs that follow, specific details of implementing each step of the simulation are described.

Data Generation

It is convenient to think of the data generation process as comprising two stages: generation of abilities and background data, and generation of item response strings. Recall that we actually conceive ability as being two–dimensional; the first dimension represents the ability assessed by Type I items, and the second dimension is the additional capability required to complete Type II items successfully. Values for the first dimension of ability (denoted θ_1) were sampled from the plausible values for 17–year-olds in the 1994 NAEP reading assessment or 1992 NAEP math assessment. We sampled 4002 cases, along with case weights, and rescaled the values so that the weighted mean and variance were zero and one, respectively. We sampled 4002 values so that for design cells with six blocks of items we could administer each block to 667 putative individuals; for cells with four blocks of items, we omitted two of the sampled values and administered each block to 1000 individuals. The same sample of 4002 values (or the first 4000 cases of that sample) was employed in every cell. When the design called for a simulated shift in ability, we simply subtracted 0.15 from each value. The distributions were slightly negatively skewed, with some suggestion of a possible ceiling effect; this lack of symmetry was more pronounced in the reading distribution than in the mathematics distribution. Figure 1 shows the approximate shape of the distributions, although the histogram does not account for case weights. Abilities on the second dimension (θ_2) and values for the background variable were pseudo–randomly sampled from the standard normal distribution.[1] After sampling, we rescaled each distribution to have exactly zero mean and unit variance. We then achieved the desired correlational structure by multiplying the matrix comprising the columns of abilities and the background variable by the Cholesky decomposition of the target correlation matrix.

We employed a modification of Bock's full–information factor analysis model (Bock, Gibbons & Muraki, 1988) to define the probability of an individual with particular values

1. Here, and throughout the simulation, pseudorandom normal numbers were generated by the polar method (Knuth, 1981; Algorithm P). Uniform numbers were generated using a custom implementation of Marsaglia's(1991) portable random number generator.

of θ_1 and θ_2 passing an item. The modifications involved two aspects. First, we employed a logistic probability model, rather than the normal ogive approach. Second, we adjusted the model to accommodate guessing. The resulting probability equation was:

$$P(x_{ij} = 1 \mid \theta_i) = g_j + \frac{1 - g_j}{1 + \exp\left[-a_j(b_j - \lambda\theta_{1i} + (1 - \lambda)\theta_{2i})\right]}$$

where *a*, *b*, and *g* are the slope, threshold, and guessing parameter of the usual three parameter logistic IRT model, λ is a mixing coefficient bounded by zero and one, and x_{ij} is equal to one when person *i* responds correctly to item *j*. Given particular values of item and person parameters, a "correct" response was generated when a uniformly distributed pseudorandom number was less than the probability derived from the equation; otherwise, a failure was generated.

Item parameters for each cell of the simulation design were selected from the values reported for the dichotomously scored items in the reading or mathematics assessments for 17–year–olds in the 1992 NAEP technical report. Type II items were chosen from those items with guessing parameters fixed at 0.0; Type I items were chosen from among the others. We made an effort to keep the average threshold parameters in each block of six items (cells 11–20) or 12 items (cells 1–10) near the overall mean threshold of approximately negative 0.5 for reading and 0.0 for mathematics. When cells differed only in the number of common versus not common Type I items, the same generating item parameters were used whenever possible. Within a cell, the same item parameters were used to generate wave one and wave two data.

Calibration and Equating

The wave one data were calibrated using Bilog 3 for DOS, with strong priors constraining the intercept parameters of Type II items to be near zero (a value of 0.001 was actually employed to avoid possible numerical difficulties associated with fixing a prior mean that fell on the boundary of values allowed under a beta distribution). A special computer program automatically generated the Bilog command file. At the completion of Bilog's item parameter estimation, estimates were preserved in a copy of the item output file; we then used Bilog's expected *a posteriori* ability estimation module, rescaling so that the sample ability estimates had a mean of zero and unit variance. We generated wave two data, and another special program generated a new Bilog command file that placed strong priors on the common items, fixing them at the values output at the completion of phase two in the previous estimation. The program also read the relevant Bilog output file from wave one estimation to find the rescaling constants that were employed to achieve standardized ability estimates, and wrote the new Bilog command file in such a way that rescaling of the new ability estimates would employ the same constants. The resultant ability estimates thus represent estimates equated by the IRT invariance strategy. The same Bilog command file was employed for invariance equating of the wave two data with an ability shift (on the θ_1 dimension only) of 0.15 standard deviations.

We implemented common item equating by a similar mechanism. A special computer program generated a Bilog command file that simultaneously scaled all 8000 or 8004 cases (including both wave one and wave two data). We treated common items as the same regardless of which time they were employed; non–common items were treated as distinct, even though they had the same generating values at both times. Thus, in a cell with 48 items of which 24 were common, Bilog was instructed to scale 72 items, divided among 12 test forms. Once again, we fixed the guessing parameters of Type II items at approximately zero. We instructed Bilog to produce ability estimates in the standard metric. Then a separate program derived rescaling constants A_{CI} and B_{CI} based on the first 4000 or 4002 estimated abilities from the common scaling run. The program wrote a Bilog command file that fixed all item parameters at their previously estimated values, and applied the newly derived scaling constants. The resultant abilities thus represent estimates equated by the common item strategy.

We implemented multiple group IRT by jointly calibrated wave one and wave two data together (treated as two groups) using Bilog–MG. The scales were linked by virtue of the joint estimation of item parameters. This implementation illustrates the potential of multiple group IRT if it were used to simultaneously scale two or more waves of trend data. Such a use would be possible when a new trendline was established or an old one rescaled to improve comparability across years.

Generation and Assessment of Plausible Values

The problem of generating plausible values was considerably simpler in our case than in the real NAEP analyses, since the multivariate nature of the background variables was simplified to a univariate relationship. Recall that at the data generation stage, we produced a background variable, γ, which was correlated with θ_1 and θ_2; the value of the correlation was γ. The background variable was scaled to have mean zero and variance one, or a mean of $-0.15/\gamma$ when the ability was shifted. If ability were truly unidimensional, then the mean and variance of the posterior distribution of a particular person's θ could be obtained by appropriate manipulations of the integral of θ times its posterior density, and θ^2 times the posterior density of θ. The posterior density is proportional to:

$$P(\theta_i \mid x_i, y_i, \gamma, \sigma) = \phi((\theta_i - \gamma y_i)/\sigma)/\sigma \prod_j p(x_{ij} \mid \theta_i),$$

where x_{ij} is the j^{th} element of individual *i*'s item response string, σ is the residual standard deviation in the regression (and is thus wholly determined by γ, since it is equal to the square root of $1 - \gamma^2$), and $\phi(z)$ denotes the standard normal probability density function evaluated at *z*. We evaluated these integrals numerically for each individual's ability, following a procedure that involved several steps. First, we identified an appropriate range for integration (i.e., a range over which the function was numerically non–zero). Next, we integrated to get the normalizing constant. Finally, we integrated θ and θ^2 times the normalized posterior density. The posterior mean was then taken to be the numerical

result of the integral involving θ, and the posterior variance was the second integral minus the squared posterior mean. We then generated five plausible values for each individual, by randomly sampling from the normal distribution with the obtained mean and variance. We calculated the weighted means, variances, skew indexes, and kurtosis indexes, as well as nine quantiles, for the 4000 (or 4002) replications of each of the five plausible values. The results, presented in Section A and discussed in the *Results* section, are the means of the five instances of each statistic.

Results

The results of this study suggest that the common item equating and scale linking used in NAEP perform rather well on the average, even when each student takes only one quarter of the items on the scale and the equating is based disproportionately on one type of item. The average bias due to equating is the estimated difference in the mean of the scaled ability distribution between one assessment wave and the next minus the change in the actual means of the distributions of ability parameters for the two waves. While the bias was statistically reliable in some cases (it was several times its standard error) it was never large in comparison to the real changes that have been observed in NAEP. The maximum bias in the scale mean under any of the conditions examined was only about 0.01 standard deviations in reading and 0.02 standard deviations in mathematics, which (given a typical NAEP scale standard deviation of about 40) is about 0.5 to 1.0 scale points. Multiple group IRT methods have the potential to produce even smaller biases. The results for simulations based on ability distributions and item parameters for reading and mathematics are discussed in detail below, followed by those of the simulation of multiple group IRT equating for both subject matters.

Reading

For common item equations, the maximum bias in the scale mean under any of the conditions examined was only about 0.011 standard deviations which (given a typical NAEP scale standard deviation of about 40) is about 0.5 scale points. Table A.1 presents the mean bias (mean for wave 2 minus mean for wave 1 minus the true change) for 80 conditions selected from the design which are the conditions under which it is most difficult to achieve equating and scale linking.

The pattern of bias suggests a few generalizations. Common item equating generally appears to work best when the proportion of Type I items on the scale is the same as the proportion of Type I items used as common items for equating. When these proportions are highly unequal (that is when the common items used for equating are disproportionately Type I items but the entire scale is not), then equating is poorest. Population shifts generally, though not always, make equating more difficult. Surprisingly, these data suggest that scale linking is not necessarily less biased for longer scales or when

more of the items are taken by each student. The largest bias occurred when 24 items (one half) of a 48 items scale were taken by each student.

While common item equating and scale linking performs remarkably well, it should be noted that IRT invariance equating and scale linking does not. Table A.1 shows that when the mean bias is large, the bias using IRT invariance linking can be several times as great as that of common item linking. The biases found here could be larger than 1.5 NAEP scale points, which is not negligible in absolute terms or in comparison to typical NAEP sampling standard errors.

Higher scale moments. In addition to comparing the means of the equated scales, the variances of the wave 1 and wave 2 (linked) scales were also compared. Table A.2 presents the ratio of each wave 2 scale variance to that of the original (wave 1) scale. While it appears that the scales linked by the common item equating usually had larger variances than the original scale, the increase in variance is small. Scale variances for IRT invariance linked scales appear to be somewhat closer to the original scale variances. No particularly notable patterns in the variance ratios are apparent.

The third and fourth moments of the original (wave 1) and linked distributions were also compared. The differences between these statistics for the wave 1 distribution and those of the linked (wave 2) distributions are given in tables A.3 and A.4. Since the nature of the population shift from wave 1 to wave 2 is a constant movement, one would expect these differences to be zero if the linking were perfect. It appears from these statistics that common item linking performs well, even in situations where it would be expected to be perform least well.

Comparisons of scale quantiles. Another way the linked (wave 2) distributions were compared with the original (wave 1) distributions was by comparing the quantiles (the 1 percent, 5 percent, 10 percent, 25 percent, 50 percent, 75 percent, 90 percent, 95 percent, and 99 percent points of the distribution). Figures B.2 through B.20 in Section B use these quantiles to illustrate the cumulative distribution of the original (wave 1) distribution and the four linked (wave 2) distributions for the 20 configurations of items discussed above. In each case there are two groups of ogives, with the curves in each group virtually indistinguishable from one another. One group including the original (wave 1) distribution and the linked distributions with no population change. The other group corresponds to the two linked distributions with the -0.15 population change (see figures B.2–B.21 in Section B).

These figures illustrate that the distributions match reasonably well in many cells. However, in some cells, there are differences between the quantiles of the linked distributions and what might be expected with perfect equating. These differences are often, although not always, larger in the lower quantiles than at the upper part of the distribution. Further, the differences occur even when there are no population changes. For example in cells 6, 7, and 8 (where equating is generally poorest) the 5 percent, 10 percent, and 25 percent points in the wave 2 distribution obtained by common item

linking differ from the corresponding quantiles of the wave 1 distribution by about 0.05, 0.04, and 0.03 standard deviations, respectively. These biases are statistically reliable, being several times their standard errors. Assuming a typical NAEP standard deviation of about 40 points, these biases would suggest that changes at these quantiles could be misestimated by as much as 2 NAEP scale points, which would not be negligible.

Detailed information on comparisons of scale moments. Tables A.5 and A.6 provide a more detailed report of the scale means for the 80 conditions previously discussed, including the standard errors of each mean. Note that the means of the original (wave 1) distribution are not identically zero. The reason is that, although the distribution of generating values may have had a mean of 0 and a variance of 1, the ability values estimated after scaling with a different set of items would no longer have a mean of zero. Since different cells in the design called for items with somewhat different characteristics, the means in the wave 1 distribution are slightly different in each cell of the design.

Tables A.7 and A.8 provide detailed information for the scale variances. Note that the variances of the original (wave 1) distribution of abilities are not all 1. As in the case of the means, even if the generating distribution of abilities had a mean of one, the ability values estimated after scaling with a different set of items would no longer have a variance of one. Since different cells in the design called for items with somewhat different characteristics, the variance of the wave 1 distribution is slightly different in each cell of the design.

Tables A.9 and A.10 give the corresponding values for the scale skewness, while tables A.11 and A.12 provide a summary of the scale kurtosis values. Tables of the quantiles are not included, but they have been produced and are available on request.

Effects of increasing multidimensionality. The value of λ used for Type II items in the main simulation ($\lambda = 0.9$) was chosen as probably reasonable after some examination of the literature and discussion with the members of the NAEP Validity Studies Panel. To see whether a smaller value of λ, corresponding to a higher degree of multidimensionality, would have a more deleterious effect on equating, one cell of the design was rerun with $\lambda = 0.7$, a value we considered to be too small to be realistic. In this simulation, the change of -0.15 units in ability for Type I items is accompanied by a change of only -0.105 for Type II items.

The results of this simulation, given in table A.13, suggest that even under these conditions common item linking performed about as well as under the other conditions studied. The mean bias for common item linking was -0.009 standard deviations when there was no population change and -0.011 standard deviations when there was a population change.

The effect was greater for IRT invariance linking than for common item linking, but not substantially greater than it was for IRT invariance linking with no population change.

The mean bias for invariance linking was -0.003 standard deviations when there was no population change and -0.039 standard deviations when there was a population change.

The effects of increased multidimensionality on the extreme quantiles of the distribution are considerably larger than at the mean, but not substantially larger than for the cases with less multidimensionality. Figure B.21 uses the quantiles to illustrate the cumulative distribution of the original (wave 1) distribution and the four linked (wave 2) distributions discussed in this section.

Effects of increasing the precision of the simulated values. The results previously reported are based on 10 replications of each condition. However, since the number of students is relatively large (4,000 per wave) and the generating ability distribution is identical in each replication, the variation across replications is rather small. To investigate whether results would change substantially with a larger number of replications, 50 replications of cells 3, 5, 6, 7, 8, and 9 (where equating showed the poorest performance) were run and compared with the results of the first 10. There was no substantial change in results for the first four moments or the quantiles of the distributions; in most cases, the quantiles shifted slightly only in the third decimal place.

Mathematics

The scale linking biases tended to be somewhat larger for mathematics than the corresponding biases for reading. We believe that this may be associated with the lower correlation between background variables and the mathematics ability scales. For common item equating, the maximum bias in the scale mean under any of the conditions examined was only about 0.021 standard deviations which (given a typical NAEP scale standard deviation of about 40) is about 1.0 scale points. The maximum bias using strict IRT invariance equating was more than twice as large as that for common item equating. Table A.14 presents the mean bias (mean for wave 2 minus mean for wave 1 minus the true change) for the 80 conditions selected from the design which were also examined for the reading simulation. Because the reading study suggested that IRT invariance equating was markedly inferior to common items equating, results for IRT invariance are presented only for the case of no change in population ability (as a check on previous results).

The pattern of bias suggests a few generalizations. Common item equating generally appears to work best when the proportion of Type I items on the scale is the same as the proportion of Type I items used as common items for equating. When these proportions are highly unequal (that is when the common items used for equating are disproportionately Type I items but the entire scale is not), then equating is poorest. Population shifts generally, though not always, make equating more difficult. These data also suggest that scale linking is not necessarily any more or less biased for longer scales or when more of the items are taken by each student. The two cells with the largest bias in common item equating occurred when 12 items (one half) of a 24 item scale were taken

by each student, but the next largest biases occurred when students took 24 items (one half) of a 48 item scale.

Common item equating and scale linking performs remarkably well, but IRT invariance equating and scale linking does not, even when there is no change in the population mean. Table A.14 shows that the mean bias using IRT invariance linking can be several times as great as that of common item linking. The biases found here could be almost 2.5 NAEP scale points, which is not negligible in absolute terms or in comparison to typical NAEP sampling standard errors. However we detected some problems in determining convergence for IRT invariance linking in cells 16–20, which suggests that the magnitude of these biases may be somewhat over estimated.

Higher scale moments. In addition to comparing the means of the equated scales, the variances of the wave 1 and wave 2 (linked) scales were also compared. Table A.15 presents the ratio of each wave 2 scale variance to that of the original (wave 1) scale. While it appears that the scales linked by the common item equating usually had larger variances than the original scale, the increase in variance is small (less than 6 percent). Scale variances for IRT invariance linked scales appear to be somewhat closer to the original scale variances, except in the case of cells 16–20 (where each person took half the items on a short scale).

The third and fourth moments of the original (wave 1) and linked distributions were also compared. The differences between these statistics for the wave 1 distribution and those of the linked (wave 2) distributions are given in tables A.16 and A.17. Since the nature of the population shift from wave 1 to wave 2 is a constant movement, one would expect these differences to be zero if the linking were perfect. It appears from these statistics that common item linking performs well, even in situations where it would be expected to perform least well.

Comparisons of scale quantiles. Another way the linked (wave 2) distributions were compared with the original (wave 1) distributions was by comparing the quantiles (the 1 percent, 5 percent, 10 percent, 25 percent, 50 percent, 75 percent, 90 percent, 95 percent, and 99 percent points of the distribution). Figures B.23 through B.42 use these quantiles to illustrate the cumulative distribution of the original (wave 1) distribution and the three linked (wave 2) distributions for the 20 configurations of items discussed above. In each case there are two groups of ogives, with the curves in each group very similar to one another. One group including the original (wave 1) distribution and the linked distributions with no population change. The other group corresponds to the linked distribution with the -0.15 population change.

These figures illustrate that the distributions match reasonably well in many cells. However, in some situations (such as those of cells 16, 17, and 18), there are differences between the quantile of the linked distributions and what might be expected with perfect equating. These differences are often, although not always, larger in the lower quantiles than at the upper part of the distribution. In these situations IRT invariance equating

performed particularly poorly, producing biases larger than 0.1 standard deviation at the 5 percent point. Assuming a typical NAEP standard deviation of about 40 points, these biases would suggest that changes at these quantiles could be misestimated by more than 5 NAEP scale points. These differences occur even when there are no population changes.

Common item equating usually performed substantially better than IRT invariance equating, however, some of the biases are still significant at the extremes. For example in cells 16, 17, and 18 (where equating is generally poorest) the 5 percent points in the wave 2 distribution obtained by common item linking differ from the corresponding quantiles of the wave 1 distribution by about 0.05, 0.05, and 0.04 standard deviations, respectively. These biases are statistically reliable, being several times their standard errors. Assuming a typical NAEP standard deviation of about 40 points, these biases would suggest that changes at these quantiles could be misestimated by as much as 3 NAEP scale points, which would not be negligible.

Detailed information on comparisons of scale moments. Tables A.18 and A.19 provide a more detailed report of the scale means for the 80 conditions previously discussed, including the standard errors of each mean. Note that the means of the original (wave 1) distribution are not identically zero. The reason is that, although the distribution of generating values may have had a mean of 0 and a variance of 1, the ability values estimated after scaling with a different set of items would no longer have a mean of zero. Since different cells in the design called for items with somewhat different characteristics, the means in the wave 1 distribution are slightly different in each cell of the design.

Tables A.20 and A.21 provide detailed information for the scale variances. Note that the variances of the original (wave 1) distribution of abilities are not all one. As in the case of the means, even if the generating distribution of abilities had a mean of 1, the ability values estimated after scaling with a different set of items would no longer have a variance of 1. Since different cells in the design called for items with somewhat different characteristics, the variance of the wave 1 distribution is slightly different in each cell of the design.

Tables A.22 and A.23 give the corresponding values for the scale skewness, while tables A.24 and A.25 provide a summary of the scale kurtosis values. Tables of the quantiles are not included, but they have been produced and are available on request.

Multiple Group IRT

Multiple group IRT performed extraordinarily well, even better than the common item equating procedures we studied. The maximum bias in the scale mean under any of the conditions examined was less than 0.01 standard deviations which (given a typical NAEP scale standard deviation of about 40) is about 0.5 scale points, or half of that of common item equating. In most cases, the bias was so small as to be negligible in both the mathematics and the reading simulations. Table A.26 presents the mean bias (mean for

wave 2 minus mean for wave 1 minus the true change) for 80 conditions selected from the design which are the conditions under which it is the most difficult to achieve equating and scale linking.

Higher scale moments. In addition to comparing the means of the equated scales, the variances of the wave 1 and wave 2 (linked) scales were also compared. Table A.27 presents the ratio of each wave 2 scale variance to that of the original (wave 1) scale. While it appears that the scales linked by the multiple group equating usually had smaller variances than the original scale when there was no change and usually had larger variances than the original scale when there was a change, the difference in variance is small (less than 2 percent).

The third and fourth moments of the original (wave 1) and linked distributions were also compared. The differences between these statistics for the wave 1 distribution and those of the linked (wave 2) distributions are given in tables A.28 and A.29. Since the nature of the population shift from wave 1 to wave 2 is a constant movement one would expect these differences to be zero if the linking were perfect. It appears from these statistics that multiple group linking performs well, even in situations where it would be expected to be perform least well.

Comparisons of scale quantiles. Another way the linked (wave 2) distributions were compared with the original (wave 1) distributions was by comparing the quantiles (the 1 percent, 5 percent, 10 percent, 25 percent, 50 percent, 75 percent, 90 percent, 95 percent, and 99 percent points of the distribution). The distributions match extraordinarily well in all cells, and plots of the cumulative distribution of linked distributions are indistinguishable. The differences between the quantiles of the linked distributions are generally different by no more than might be expected due to sampling error if there were perfect equating. Unlike the common item equating methods studied, these differences are no larger in the lower quantiles than at the upper part of the distribution.

Discussion

This study suggests that the common item equating and scale linking currently used in NAEP perform rather well, even when the number of common items is small, each student takes only 25 percent of the items on a scale, the ability scale is slightly multidimensional, and there are changes in the ability distribution. The bias in estimating mean performance introduced by common item equating appears to be no more than about 0.01 to 0.02 standard deviations or one–half to one point on the NAEP scale. This is small, but not entirely negligible in comparison to the sampling standard error at the mean for the nation as a whole. Explorations of the effect of increasing multidimensionality somewhat do not produce substantially larger equating bias.

It is important to recall that NAEP may create several subscales for a given subject area that are averaged to obtain an overall scale for that subject. It is tempting to believe that the biases in the subscales would cancel out and that the overall scale would be less biased than the subscales from which it is composed. This need not be the case. Consequently, the biases in the overall scale may not be less than that of the subscales. Indeed, they could be larger in comparison to the decreased standard error of the overall scale.

To apply these results to operational NAEP one might examine the equating of the 1992 to 1994 short–term trend scales in reading. There were three scales: reading for information, reading to perform a task, and reading for literary purposes. The information scale had 40 items, half of which were constructed response, but only 13 percent of the constructed response items were used as common items for equating. Therefore, the situation for the information scale most resembles cells 1 or 6. The literary and task scales had 20 and 27 items respectively, of which 65 percent and 59 percent were constructed response, but only 50 percent and 45 percent of the constructed response items were used for equating. The situation for these two scales resembles cells 13 or 18. This analogy suggests that the bias should be between 0.001 and 0.011 standard deviations (0.0 to 0.5 NAEP scale points) for the information scale and 0.003 and 0.006 standard deviations (0.1 to 0.3 NAEP scale points) for the other two scales.

It might be advisable to consider equating as introducing as much as 0.5 to 1.0 points of bias in trend comparisons. Thus a viable procedure might be to test for differences between assessment waves by testing whether the difference is greater than 1.0 scale units (the maximum equating bias found here). Alternatively, one might increase the sampling standard error by a fraction that would accomplish approximately the same result as a way to characterize the contribution of equating bias to uncertainty. That is, one might treat equating error as a fixed component in the variance of the difference between assessment wave means. Assuming approximately equal sample sizes in each assessment wave, this leads to a standard error for the difference of the form:

$$SE_{Difference} = \sqrt{SE_1^2 + SE_2^2 + \sigma^2/2}$$

where δ is the equating bias (e.g., 0.5).

Results for scale quantiles suggest more caution. While there was generally only small bias in the scale quantiles due to linking, in some cases the bias in the quantiles was substantial, up to 5 times that of the mean. This suggests that scale linking can pose problems for inferences about changes in the extremes of the distribution or about groups whose scores tend to be extreme. In the cases where linking was poorest, the 5 percent, 10 percent, and even 25 percent points shifted by an amount equivalent to as much as 2 NAEP scale points. This may be particularly important if the performance of disadvantaged groups, who tend to score substantially below the mean, continues to be an important national policy interest.

This simulation did not address the question of multiple linkings over more than two waves of data collection. While it would be naive to assume that worst–case biases would simply compound over years, it is not clear exactly how much biases increase after linking of several waves of assessments. On the other hand it seems realistic to assume that there is some compounding, and that the effects on extreme quantiles could be substantial. Even the effect on means could be nonnegligible after, say, five waves of data collection. For example, a bias in one direction of 0.005 standard deviations compounding over five waves of assessment could become a total bias of 0.025 standard deviations or about a scale point. Recalling that the bias at extreme quantiles could be five times as large, such compounding could correspond to a bias of several points at the quantiles.

Multiple group IRT models have great promise as alternatives to equating and linking based on single group IRT methods. When a multiple group model was used in this simulation, nearly all of the bias was eliminated, and the linked distribution was virtually indistinguishable from what would have been expected if there were perfect equating. The multiple group method is useful when two waves of data can be scaled together (for example, when an entire trend series is computed at once), and the advantages should be even greater when more than two waves of data are linked. On the other hand, multiple group equating would not have much advantage over conventional IRT methods if, for example, the first wave of the data was scaled separately and multiple group methods could be applied only to the second wave of the data, since the equating would not be provided internally by the multiple group model.

Trend reporting in NAEP has not, up until now, involved revisions to previous reports. However it is possible to introduce such revisions. Social statistics of many kinds are revised from time to time, and even values of fundamental physical constants are subject to periodic redetermination that alters their values. The revision of scores that occurs in multiple group IRT is a consequence of additional information (the second wave of data) which increases the precision of estimates of the scores in the first wave of data. There is a revealing parallel in the determination of values of fundamental physical constants. Experiments which estimate these constants must rely on data from other experiments which measure related constants or the relations among constants. The value of a constant may need to be revised when better data on related constants is obtained. We regard the revision of first wave of scores in multiple group IRT as logically equivalent to the revision of the value of a physical constant given new data on a related constant. While retrospective redetermination of individual test scores might pose problems, individual test scores are not provided by NAEP or other assessment programs that focus on population distributions. The merits of less biased measurements may outweigh the problems caused by slight adjustments to scores, particularly in long trend lines where equating and linking errors are likely to be greatest.

Finally, we noted that our simulations were sensitive to the background variable used in the conditioning process. We believe that most of the differences between mathematics and reading that we observed were a consequence of the fact that the correlation between the background variables and mathematics ability was lower than the correlation between

the background variables and reading ($R^2 = 0.25$ versus $R^2 = 0.52$). It is clear that changes in the background variables or their relation to achievement can affect the ability distribution generated through multiple imputation. Our simulations relied on a correlation structure with the background variables that did not change over time. That structure may not remain constant if background variables or the process by which they are collected are changed. The maintenance of a constant set of background variables for conditioning of (short– or long–term) trend data is a consideration that should not be overlooked in the operation of NAEP.

Recommendations

This research suggests some practical recommendations for practice in NAEP and other large scale assessments using NAEP–like procedures.

1. Even in the most difficult conditions usually encountered, the common item equating and scale linking procedures currently used in operational NAEP appear to introduce relatively little bias (less than one NAEP scale point) in comparisons of the means of two waves of data. There should be little bias also in comparisons of subgroup means that are relatively near the center of the overall populations. The fact that these procedures are also straightforward and well understood supports their continuation.

2. The common item equating and scale linking procedures currently used in operational NAEP introduce substantially more bias (up to two NAEP scale points) in comparisons of the extreme percentiles of two populations. We recommend caution in comparisons of extreme percentiles over time or comparisons over time of the means of population subgroups which differ substantially from the overall population mean. Such cautions would apply also to examination of trends over time in proportions of the population at extremely high achievement levels. In these cases, the sampling standard errors may substantially understate the true uncertainties of trends. In such cases the use of a conservative test that the scale difference is larger than some nonzero value (e.g., 2 NAEP scale points) may be warranted as a test of the null hypothesis of no trend.

3. Strict IRT invariance equating and scale linking should not be used in NAEP or other large scale assessments. It introduces substantially more bias than the procedures currently used in NAEP.

4. Multiple group IRT methods have considerable scientific merit for equating and scale linking. These methods have the potential of practically eliminating bias in scale linking, even in the situations where current methods are weakest. When all waves of data can be analyzed together,

multiple group IRT has no apparent disadvantages. When all waves of data cannot be scaled together (as in NAEP trend reporting), multiple group IRT methods have the disadvantage that the linking of a second (or later) wave of data alters scores on the first wave of data. We believe that this is not a fatal flaw. Social statistics of many kinds are revised from time to time and even values of fundamental physical constants are subject to periodic redetermination that alters their values. The revision of scores that occurs in multiple group IRT is a consequence of additional information (the second wave of data) which increases the precision of estimates of the scores in the first wave of data. While retrospective redetermination of individual test scores might pose problems, individual test scores are not provided by NAEP or other assessment programs that focus on population distributions. We believe that the merits of less biased measurements may outweigh the problems caused by slight adjustments to scores.

5. Although current NAEP procedures appear adequate for comparisons of population means across two or three waves of data, they do not ensure that equating and linking biases will not compromise long trend lines and particularly trends of extreme percentiles. Therefore, the data underlying long trend lines should be periodically reanalyzed using methods, such as multiple group IRT, which can minimize equating and linking bias.

Section A: Tables

Table A.1 Average Scale Linking Bias (Reading Simulation)

Cell	Number of: Common Items	Type I Items	Common Type I Items	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Invariance Equating	-0.15 Population Change: Common Item Equating
48 Total Items, 12 Items Taken							
1	6	6	1	-0.001	-0.006	0.010	0.002
2	6	6	3	-0.008	-0.002	0.011	0.002
3	6	7	1	-0.000	-0.001	0.021	0.003
4	6	8	3	-0.003	-0.001	0.018	0.006
5	12	6	6	0.022	-0.004	0.025	-0.007
48 Total Items, 24 Items Taken							
6	12	12	2	0.002	-0.006	0.032	0.011
7	12	12	6	-0.003	-0.004	0.028	0.010
8	12	14	2	-0.004	-0.010	0.037	0.007
9	12	16	6	-0.002	-0.006	0.027	0.008
10	24	12	12	0.000	0.004	0.009	-0.001
24 Total Items, 6 Items Taken							
11	3	3	1	-0.001	0.004	0.005	0.004
12	3	3	2	-0.001	0.001	-0.001	0.006
13	3	4	1	0.001	0.004	0.009	0.005
14	3	4	2	0.001	0.005	0.005	0.005
15	6	3	3	-0.002	0.002	-0.007	0.000
24 Total Items, 12 Items Taken							
16	6	6	2	0.000	-0.000	0.014	0.007
17	6	6	4	0.001	0.004	0.012	0.001
18	6	8	2	0.001	-0.003	0.016	0.003
19	6	8	4	-0.000	-0.002	0.012	0.003
20	12	6	6	-0.001	0.000	-0.002	-0.002

Note: Standard errors are typically less than 0.003.

Table A.2 Ratio of Wave 2 to Wave 1 Scale Variances (Reading Simulation)

	Number of:			No Population Change		-0.15 Population Change	
Cell	Common Items	Type I Items	Common Type I Items	Invariance Equating	Item Equating	Common Invariance Equating	Common Item Equating
48 Total Items, 12 Items Taken							
1	6	6	1	0.997	1.036	0.991	1.025
2	6	6	3	1.003	1.029	0.991	1.030
3	6	7	1	0.999	1.042	0.986	1.041
4	6	8	3	0.996	1.027	0.994	1.027
5	12	6	6	0.992	1.028	0.977	1.029
48 Total Items, 24 Items Taken							
6	12	12	2	1.003	1.055	0.988	1.054
7	12	12	6	1.000	1.052	0.990	1.051
8	12	14	2	0.996	1.048	0.984	1.051
9	12	16	6	0.998	1.055	0.997	1.055
10	24	12	12	0.992	1.054	0.984	1.055
24 Total Items, 6 Items Taken							
11	3	3	1	0.997	0.986	0.994	0.988
12	3	3	2	0.996	0.984	0.997	0.985
13	3	4	1	0.993	0.992	0.994	0.988
14	3	4	2	1.002	0.992	1.007	0.986
15	6	3	3	0.996	0.984	1.001	0.974
24 Total Items, 12 Items Taken							
16	6	6	2	0.996	1.015	0.996	1.026
17	6	6	4	1.001	1.029	1.001	1.023
18	6	8	2	1.006	1.025	0.996	1.026
19	6	8	4	0.996	1.034	0.998	1.024
20	12	6	6	0.996	1.013	0.995	1.021

Table A.3 Differences Between Wave 2 and Wave 1 Scale Skewness (Reading Simulation)

	Number of:			No Population Change		-0.15 Population Change	
Cell	Common Items	Type I Items	Common Type I Items	Invariance Equating	Item Equating	Common Invariance Equating	Common Item Equating
48 Total Items, 12 Items Taken							
1	6	6	1	0.009	0.020	0.021	-0.007
2	6	6	3	-0.001	-0.009	-0.019	-0.011
3	6	7	1	0.006	0.004	0.002	0.002
4	6	8	3	-0.004	-0.020	-0.006	-0.002
5	12	6	6	0.011	0.019	0.008	-0.008
48 Total Items, 24 Items Taken							
6	12	12	2	-0.011	-0.015	-0.018	-0.016
7	12	12	6	0.006	0.016	-0.014	0.006
8	12	14	2	0.005	0.000	0.016	0.003
9	12	16	6	0.003	0.013	-0.006	0.000
10	24	12	12	-0.017	0.003	-0.018	0.018
24 Total Items, 6 Items Taken							
11	3	3	1	-0.012	-0.021	0.002	-0.009
12	3	3	2	0.003	-0.013	-0.015	-0.011
13	3	4	1	0.001	-0.007	0.008	-0.009
14	3	4	2	0.006	-0.019	-0.018	-0.025
15	6	3	3	-0.011	-0.025	-0.008	-0.025
24 Total Items, 12 Items Taken							
16	6	6	2	0.003	-0.004	0.013	-0.000
17	6	6	4	-0.021	-0.007	-0.012	-0.018
18	6	8	2	-0.003	0.005	0.015	-0.009
19	6	8	4	-0.002	0.005	0.001	-0.006
20	12	6	6	-0.016	-0.016	-0.024	-0.020

Note: Standard errors are typically below 0.008.

Table A.4 Differences Between Wave 2 and Wave 1 Scale Kurtosis (Reading Simulation)

	Number of:			No Population Change		-0.15 Population Change	
Cell	Common Items	Type I Items	Common Type I Items	Invariance Equating	Common Item Equating	Invariance Equating	Common Item Equating
48 Total Items, 12 Items Taken							
1	6	6	1	-0.000	-0.013	0.013	-0.019
2	6	6	3	0.030	0.038	0.053	0.025
3	6	7	1	-0.023	-0.036	-0.023	-0.033
4	6	8	3	0.008	-0.036	-0.012	-0.034
5	12	6	6	0.013	-0.023	0.016	0.007
48 Total Items, 24 Items Taken							
6	12	12	2	-0.025	0.029	0.011	0.016
7	12	12	6	0.000	0.008	0.004	0.013
8	12	14	2	0.001	0.030	0.011	0.037
9	12	16	6	-0.001	0.037	0.009	0.025
10	24	12	12	0.017	0.038	0.017	0.023
24 Total Items, 6 Items Taken							
11	3	3	1	0.005	0.007	0.004	-0.006
12	3	3	2	-0.019	-0.029	-0.014	-0.028
13	3	4	1	-0.030	-0.025	0.006	-0.032
14	3	4	2	0.030	0.001	0.043	0.031
15	6	3	3	0.006	-0.004	-0.002	-0.008
24 Total Items, 12 Items Taken							
16	6	6	2	0.032	0.008	0.017	0.002
17	6	6	4	0.029	-0.003	0.010	-0.009
18	6	8	2	0.008	-0.025	0.003	0.000
19	6	8	4	0.026	-0.007	0.025	0.018
20	12	6	6	-0.002	-0.009	0.009	-0.040

Note: Standard errors are typically below 0.025.

Table A.5 48 Total Items: Scale Means (100*Standard Error) for Two Equating Methods (Reading Simulation)

Cell	Number of: Common Items	Type I Items	Common Type I Items	Wave 1	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Invariance Equating	-0.15 Population Change: Common Item Equating
12 Items Taken								
1	6	6	1	-0.011	-0.012	-0.016	-0.151	-0.159
				(0.229)	(0.311)	(0.153)	(0.179)	(0.171)
2	6	6	3	-0.008	-0.016	-0.010	-0.147	-0.156
				(0.161)	(0.191)	(0.173)	(0.210)	(0.207)
3	6	7	1	-0.010	-0.011	-0.011	-0.139	-0.157
				(0.158)	(0.197)	(0.228)	(0.193)	(0.232)
4	6	8	3	-0.010	-0.013	-0.011	-0.142	-0.154
				(0.170)	(0.325)	(0.138)	(0.151)	(0.163)
5	12	6	6	-0.006	0.016	-0.009	-0.131	-0.163
				(0.169)	(0.292)	(0.203)	(0.285)	(0.238)
24 Items Taken								
6	12	12	2	-0.006	-0.004	-0.012	-0.124	-0.145
				(0.206)	(0.154)	(0.197)	(0.256)	(0.218)
7	12	12	6	-0.006	-0.009	-0.010	-0.128	-0.146
				(0.126)	(0.251)	(0.308)	(0.171)	(0.296)
8	12	14	2	-0.001	-0.005	-0.011	-0.114	-0.143
				(0.124)	(0.327)	(0.160)	(0.142)	(0.271)
9	12	16	6	-0.001	-0.004	-0.007	-0.124	-0.143
				(0.169)	(0.192)	(0.215)	(0.350)	(0.327)
10	24	12	12	-0.009	-0.009	-0.005	-0.150	-0.160
				(0.181)	(0.110)	(0.248)	(0.210)	(0.268)

Table A.6 24 Total Items: Scale Means (100*Standard Error) for Two Equating Methods (Reading Simulation)

Cell	Number of: Common Items	Type I Items	Common Type I Items	Wave 1	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Invariance Equating	-0.15 Population Change: Common Item Equating
6 Items Taken								
11	3	3	1	-0.008 0.225)	-0.009 (0.166)	-0.004 (0.234)	-0.153 (0.220)	-0.154 (0.172)
12	3	3	2	-0.006 (0.233)	-0.007 (0.201)	-0.005 (0.113)	-0.157 (0.220)	-0.150 (0.089)
13	3	4	1	-0.008 (0.158)	-0.007 (0.220)	-0.003 (0.159)	-0.149 (0.311)	-0.152 (0.145)
14	3	4	2	-0.007 (0.179)	-0.007 (0.274)	-0.003 (0.170)	-0.152 (0.159)	-0.152 (0.140)
15	6	3	3	-0.007 (0.112)	-0.009 (0.237)	-0.005 (0.136)	-0.164 (0.145)	-0.157 (0.206)
12 Items Taken								
16	6	6	2	-0.010 (0.164)	-0.010 (0.167)	-0.010 (0.207)	-0.146 (0.242)	-0.153 (0.155)
17	6	6	4	-0.007 (0.253)	-0.006 (0.300)	-0.002 (0.246)	-0.145 (0.252)	-0.156 (0.194)
18	6	8	2	-0.005 (0.203)	-0.004 (0.334)	-0.009 (0.152)	-0.140 (0.176)	-0.152 (0.208)
19	6	8	4	-0.005 (0.278)	-0.006 (0.168)	-0.007 (0.186)	-0.143 (0.297)	-0.152 (0.261)
20	12	6	6	-0.011 (0.214)	-0.011 (0.303)	-0.011 (0.252)	-0.163 (0.210)	-0.163 (0.219)

Table A.7 48 Total Items: Scale Variances (100*Standard Error) for Two Equating Methods (Reading Simulation)

Cell	Number of: Common Items	Common Type I Items	Type I Items	Wave 1	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Invariance Equating	-0.15 Population Change: Common Item Equating
12 Items Taken								
1	6	6	1	1.096 (0.297)	1.093 (0.671)	1.135 (0.325)	1.086 (0.509)	1.124 (0.397)
2	6	6	3	1.097 (0.231)	1.100 (0.265)	1.129 (0.324)	1.087 (0.344)	1.129 (0.334)
3	6	7	1	1.100 (0.465)	1.099 (0.291)	1.146 (0.226)	1.085 (0.494)	1.146 (0.240)
4	6	8	3	1.096 (0.382)	1.091 (0.367)	1.125 (0.248)	1.090 (0.478)	1.125 (0.331)
5	12	6	6	1.101 (0.213)	1.092 (0.468)	1.132 (0.543)	1.075 (0.771)	1.133 (0.450)
24 Items Taken								
6	12	12	2	1.131 (0.254)	1.135 (0.218)	1.193 (0.478)	1.117 (0.330)	1.191 (0.345
7	12	12	6	1.136 (0.274)	1.136 (0.354)	1.195 (0.417)	1.125 (0.326)	1.194 (0.375)
8	12	14	2	1.129 (0.397)	1.124 (0.348)	1.184 (0.296)	1.112 (0.449)	1.186 (0.365)
9	12	16	6	1.132 (0.398)	1.129 (0.474)	1.194 (0.377)	1.128 (0.552)	1.193 (0.353)
10	24	12	12	1.130 (0.401)	1.121 (0.396)	1.191 (0.357)	1.111 (0.580)	1.192 (0.416)

Table A.8 24 Total Items: Scale Variances (100*Standard Error) for Two Equating Methods (Reading Simulation)

Cell	Number of: Common Items	Type I Items	Type I Items	Wave 1	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Invariance Equating	-0.15 Population Change: Common Item Equating
6 Items Taken								
11	3	3	1	1.054 (0.198)	1.051 (0.432)	1.040 (0.380)	1.048 (0.620)	1.042 (0.206)
12	3	3	2	1.056 (0.467)	1.052 (0.450)	1.039 (0.384)	1.053 (0.307)	1.040 (0.313)
13	3	4	1	1.055 (0.219)	1.048 (0.459)	1.047 (0.363)	1.048 (0.196)	1.042 (0.254)
14	3	4	2	1.050 (0.422)	1.052 (0.266)	1.041 (0.262)	1.057 (0.502)	1.035 (0.296)
15	6	3	3	1.054 (0.326)	1.050 (0.639)	1.038 (0.381)	1.056 (0.321)	1.027 (0.315)
12 Items Taken								
16	6	6	2	1.099 (0.297)	1.095 (0.314)	1.115 (0.459)	1.095 (0.284)	1.127 (0.391)
17	6	6	4	1.096 (0.210)	1.097 (0.444)	1.128 (0.390)	1.097 (0.318)	1.121 (0.422)
18	6	8	2	1.096 (0.355)	1.103 (0.501)	1.123 (0.311)	1.092 (0.320)	1.125 (0.343)
19	6	8	4	1.094 (0.404)	1.090 (0.889)	1.131 (0.334)	1.092 (1.150)	1.120 (0.259)
20	12	6	6	1.104 (0.334)	1.099 (0.402)	1.118 (0.302)	1.098 (0.541)	1.127 (0.311)

Table A.9 48 Total Items: Scale Skewness (100*Standard Error) for Two Equating Methods (Reading Simulation)

Cell	Number of: Common Items	Type I Items	Common Type I Items	Wave 1	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Invariance Equating	-0.15 Population Change: Common Item Equating
12 Items Taken								
1	6	6	1	-0.076	-0.086	-0.096	-0.097	-0.069
				(0.555)	(1.014)	(0.667)	(0.281)	(0.644)
2	6	6	3	-0.097	-0.096	-0.088	-0.077	-0.086
				(0.724)	(0.869)	(0.645)	(0.803)	(0.759)
3	6	7	1	-0.086	-0.092	-0.090	-0.087	-0.087
				(0.687)	(0.451)	(0.558)	(0.620)	(0.639)
4	6	8	3	-0.094	-0.091	-0.075	-0.088	-0.092
				(0.525)	(0.620)	(0.683)	(0.773)	(0.608)
5	12	6	6	-0.083	-0.094	-0.102	-0.092	-0.076
				(0.665)	(0.771)	(0.610)	(0.684)	(0.861)
24 Items Taken								
6	12	12	2	-0.144	-0.133	-0.129	-0.125	-0.127
				(0.488)	(0.756)	(0.947)	(0.490)	(0.840)
7	12	12	6	-0.134	-0.139	-0.150	-0.120	-0.140
				(0.497)	(0.873)	(0.596)	(0.706)	(0.828)
8	12	14	2	-0.113	-0.118	-0.113	-0.129	-0.116
				(0.767)	(0.582)	(0.504)	(1.068)	(0.733)
9	12	16	6	-0.128	-0.131	-0.141	-0.121	-0.128
				(0.560)	(0.638)	(0.522)	(0.630)	(0.580)
10	24	12	12	-0.130	-0.113	-0.133	-0.112	-0.148
				(0.774)	(0.581)	(0.769)	(0.527)	(0.313)

Table A.10 24 Total Items: Scale Skewness (100*Standard Error) for Two Equating Methods (Reading Simulation)

Cell	Number of: Common Items	Type I Items	Common Type I Items	Wave 1	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Invariance Equating	-0.15 Population Change: Common Item Equating
6 Items Taken								
11	3	3	1	-0.059	-0.047	-0.038	-0.061	-0.050
				(0.570)	(0.290)	(0.698)	(0.650)	(0.769)
12	3	3	2	-0.054	-0.057	-0.041	-0.038	-0.043
				(0.437)	(0.661)	(0.534)	(0.612)	(0.469)
13	3	4	1	-0.050	-0.051	-0.044	-0.059	-0.041
				(0.639)	(0.704)	(0.725)	(0.686)	(0.633)
14	3	4	2	-0.061	-0.068	-0.042	-0.043	-0.036
				(0.469)	(0.802)	(0.451)	(0.871)	(0.701)
15	6	3	3	-0.067	-0.056	-0.041	-0.058	-0.042
				(0.533)	(0.536)	(0.453)	(0.727)	(0.735)
12 Items Taken								
16	6	6	2	-0.081	-0.083	-0.076	-0.094	-0.080
				(0.540)	(0.779)	(0.448)	(0.737)	(0.692)
17	6	6	4	-0.087	-0.067	-0.080	-0.075	-0.070
				(0.628)	(0.468)	(0.505)	(0.659)	(0.680)
18	6	8	2	-0.084	-0.081	-0.089	-0.098	-0.075
				(0.435)	(0.405)	(0.561)	(0.494)	(0.570)
19	6	8	4	-0.081	-0.079	-0.086	-0.082	-0.075
				(0.759)	(0.800)	(0.619)	(0.981)	(0.652)
20	12	6	6	-0.096	-0.080	-0.080	-0.072	-0.075
				(0.554)	(0.620)	(0.828)	(0.620)	(0.521)

Table A.11 48 Total Items: Scale Kurtosis (100*Standard Error) for Two Equating Methods (Reading Simulation)

Cell	Number of: Common Items	Type I Items	Common Type I Items	Wave 1	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Invariance Equating	-0.15 Population Change: Common Item Equating
12 Items Taken								
1	6	6	1	-0.029 (1.282)	-0.028 (1.276)	-0.015 (1.088)	-0.042 (1.413)	-0.010 (1.032)
2	6	6	3	0.021 (1.295)	-0.009 (1.249)	-0.017 (1.475)	-0.032 (1.221)	-0.004 (1.192)
3	6	7	1	-0.067 (0.743)	-0.044 (1.544)	-0.031 (0.981)	-0.044 (1.015)	-0.034 (0.994)
4	6	8	3	-0.034 (1.050)	-0.042 (1.083)	0.002 (1.200)	-0.021 (0.837)	-0.000 (1.087)
5	12	6	6	-0.012 (1.451)	-0.025 (1.253)	0.011 (1.136)	-0.028 (1.140)	-0.020 (1.355)
24 Items Taken								
6	12	12	2	-0.029 (0.849)	-0.004 (1.314)	-0.058 (1.142)	-0.040 (1.922)	-0.044 (0.948)
7	12	12	6	-0.038 (1.464)	-0.038 (1.419)	-0.046 (1.471)	-0.042 (1.405)	-0.051 (0.984)
8	12	14	2	-0.093 (1.345)	-0.095 (2.012)	-0.123 (1.987)	-0.104 (1.732)	-0.130 (1.230)
9	12	16	6	-0.035 (0.795)	-0.034 (1.490)	-0.072 (1.424)	-0.044 (1.366)	-0.060 (1.876)
10	24	12	12	-0.025 (1.650)	-0.042 (1.220)	-0.063 (1.425)	-0.042 (1.558)	-0.048 (1.537)

Table A.12 24 Total Items: Scale Kurtosis (100*Standard Error) for Two Equating Methods (Reading Simulation)

Cell	Number of: Common Items	Type I Items	Common Type I Items	Wave 1	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Invariance Equating	-0.15 Population Change: Common Item Equating
6 Items Taken								
11	3	3	1	-0.000	-0.006	-0.007	-0.004	0.006
				(1.350)	(1.130)	(1.635)	(0.679)	(1.234)
12	3	3	2	-0.030	-0.011	-0.002	-0.016	-0.002
				(0.928)	(1.118)	(1.147)	(0.811)	(0.817)
13	3	4	1	-0.019	0.011	0.007	-0.025	0.013
				(1.114)	(1.411)	(1.508)	(1.322)	(1.434)
14	3	4	2	0.014	-0.016	0.014	-0.029	-0.017
				(1.176)	(0.984)	(1.334)	(1.305)	(1.036)
15	6	3	3	-0.002	-0.008	0.003	0.000	0.007
				(1.172)	(0.894)	(0.956)	(1.529)	(0.564)
12 Items Taken								
16	6	6	2	-0.006	-0.038	-0.013	-0.022	-0.007
				(1.480)	(1.462)	(1.373)	(0.937)	(0.789)
17	6	6	4	-0.015	-0.045	-0.012	-0.026	-0.007
				(1.250)	(1.057)	(0.732)	(1.298)	(1.808)
18	6	8	2	-0.022	-0.029	0.003	-0.024	-0.022
				(1.070)	(0.918)	(1.229)	(1.343)	(0.940)
19	6	8	4	-0.014	-0.040	-0.007	-0.039	-0.032
				(0.978)	(1.127)	(1.142)	(1.589)	(1.317)
20	12	6	6	-0.022	-0.020	-0.013	-0.030	0.018
				(0.813)	(0.953)	(1.488)	(1.251)	(1.455)

Table A.13 Moments and Quantiles for Cell 8 (Reading Simulation) with λ = 0.7 (100*Standard Error)

	No Population Change			-0.15 Population Change	
	Wave 1	Invariance Equating	Common Item Equating	Invariance Equating	Common Item Equating
Moments					
Mean	-0.005 (0.121)	-0.008 (0.245)	-0.014 (0.165)	-0.116 (0.154)	-0.144 (0.194)
Variance	1.141 (0.278)	1.146 (0.403)	1.212 (0.212)	1.129 (0.306)	1.202 (0.222)
Skew	-0.100 (1.116)	-0.100 (0.849)	-0.117 (0.639)	-0.106 (0.523)	-0.105 (0.749)
Kurtosis	-0.065 (1.458)	-0.075 (0.877)	-0.099 (1.170)	-0.072 (1.650)	-0.115 (1.028)
Quantiles					
1%	-2.541 (1.089)	-2.549 (0.858)	-2.639 (0.970)	-2.648 (1.108)	-2.746 (1.204)
5%	-1.805 (0.633)	-1.807 (0.510)	-1.872 (0.565)	-1.901 (0.382)	-1.988 (0.601)
10%	-1.406 (0.501)	-1.411 (0.428)	-1.459 (0.406)	-1.505 (0.413)	-1.576 (0.326)
25%	-0.723 (0.238)	-0.731 (0.432)	-0.753 (0.367)	-0.829 (0.180)	-0.888 (0.315)
50%	0.020 (0.350)	0.014 (0.419)	0.016 (0.247)	-0.097 (0.283)	-0.118 (0.315)
75%	0.731 (0.250)	0.729 (0.296)	0.746 (0.316)	0.617 (0.269)	0.616 (0.295)
90%	1.350 (0.268)	1.353 (0.396)	1.382 (0.303)	1.233 (0.442)	1.252 (0.484)
95%	1.713 (0.602)	1.713 (0.460)	1.753 (0.392)	1.597 (0.269)	1.620 (0.586)
99%	2.387 (1.133)	2.373 (1.080)	2.417 (0.937)	2.246 (0.551)	2.283 (0.660)

Table A.14 Average Scale Linking Bias (Mathematics Simulation)

Cell	Number of: Common Items	Type I Items	Common Type I Items	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Common Item Equating
48 Total Items, 12 Items Taken						
1	6	6	1	-0.003	-0.013	-0.004
2	6	6	3	0.001	-0.004	0.005
3	6	7	1	-0.004	-0.012	0.006
4	6	8	3	-0.007	-0.006	0.006
5	12	6	6	0.000	-0.001	-0.008
48 Total Items, 24 Items Taken						
6	12	12	2	-0.004	-0.016	0.003
7	12	12	6	-0.000	-0.010	0.008
8	12	14	2	0.000	-0.013	0.001
9	12	16	6	-0.001	-0.011	0.006
10	24	12	12	0.002	-0.006	-0.007
24 Total Items, 6 Items Taken						
11	3	3	1	0.001	0.000	-0.002
12	3	3	2	0.001	0.003	0.010
13	3	4	1	0.002	-0.002	0.001
14	3	4	2	-0.004	-0.005	0.001
15	6	3	3	0.003	0.007	0.003
24 Total Items, 12 Items Taken						
16	6	6	2	0.017	-0.009	0.008
17	6	6	4	0.031	-0.004	0.021
18	6	8	2	0.021	-0.012	-0.001
19	6	8	4	0.049	-0.004	0.019
20	12	6	6	0.022	0.005	-0.003

Note: Standard errors are typically less than 0.003.

Table A.15 Ratio of Wave 2 to Wave 1 Scale Variances (Mathematics Simulation)

Cell	Number of: Common Items	Type I Items	Common Type I Items	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Common Item Equating
48 Total Items, 12 Items Taken						
1	6	6	1	1.002	1.050	1.042
2	6	6	3	0.991	1.037	1.037
3	6	7	1	0.998	1.046	1.038
4	6	8	3	0.986	1.043	1.036
5	12	6	6	0.985	1.040	1.029
48 Total Items, 24 Items Taken						
6	12	12	2	0.997	1.028	1.017
7	12	12	6	0.999	1.026	1.022
8	12	14	2	0.999	1.027	1.017
9	12	16	6	0.995	1.018	1.027
10	24	12	12	0.993	1.034	1.019
24 Total Items, 6 Items Taken						
11	3	3	1	0.994	1.017	1.015
12	3	3	2	1.015	1.027	1.030
13	3	4	1	0.977	1.019	1.003
14	3	4	2	1.002	1.028	1.018
15	6	3	3	1.002	0.975	0.969
24 Total Items, 12 Items Taken						
16	6	6	2	0.864	1.053	1.037
17	6	6	4	0.893	1.058	1.049
18	6	8	2	0.799	1.051	1.051
19	6	8	4	0.812	1.043	1.039
20	12	6	6	0.896	1.046	1.041

Table A.16 Differences Between Wave 2 and Wave 1 Scale Skewness (Mathematics Simulation)

	Number of:			No Population Change		-0.15 Population Change
Cell	Common Items	Type I Items	Common Type I Items	Invariance Equating	Common Item Equating	Common Item Equating
48 Total Items, 12 Items Taken						
1	6	6	1	0.001	-0.024	-0.031
2	6	6	3	-0.010	-0.011	-0.023
3	6	7	1	-0.001	-0.015	-0.028
4	6	8	3	-0.002	-0.016	-0.025
5	12	6	6	-0.008	-0.036	-0.030
48 Total Items, 24 Items Taken						
6	12	12	2	-0.014	-0.045	-0.042
7	12	12	6	-0.006	-0.028	-0.041
8	12	14	2	-0.012	-0.040	-0.035
9	12	16	6	0.002	-0.031	-0.030
10	24	12	12	-0.007	-0.010	-0.015
24 Total Items, 6 Items Taken						
11	3	3	1	-0.015	0.004	0.003
12	3	3	2	0.001	0.004	-0.005
13	3	4	1	0.024	0.022	0.030
14	3	4	2	0.001	0.008	0.002
15	6	3	3	-0.014	-0.002	0.013
24 Total Items, 12 Items Taken						
16	6	6	2	0.059	-0.017	-0.021
17	6	6	4	0.070	-0.018	-0.008
18	6	8	2	0.080	-0.014	-0.039
19	6	8	4	0.141	-0.005	-0.019
20	12	6	6	0.068	-0.022	-0.031

Note: Standard errors are typically less than 0.008

Table A.17 Differences Between Wave 2 and Wave 1 Scale Kurtosis (Mathematics Simulation)

	Number of:			No Population Change		-0.15 Population Change
Cell	Common Items	Type I Items	Common Type I Items	Invariance Equating	Common Item Equating	Common Item Equating
48 Total Items, 12 Items Taken						
1	6	6	1	0.011	0.004	-0.008
2	6	6	3	-0.018	0.001	-0.035
3	6	7	1	-0.014	-0.012	-0.023
4	6	8	3	-0.018	-0.039	-0.022
5	12	6	6	-0.021	-0.024	-0.042
48 Total Items, 24 Items Taken						
6	12	12	2	0.022	0.004	-0.023
7	12	12	6	0.024	0.024	-0.009
8	12	14	2	-0.010	-0.006	-0.023
9	12	16	6	-0.026	-0.016	-0.034
10	24	12	12	-0.017	0.016	0.000
24 Total Items, 6 Items Taken						
11	3	3	1	-0.016	-0.028	-0.071
12	3	3	2	0.019	-0.045	-0.046
13	3	4	1	-0.001	-0.019	-0.045
14	3	4	2	-0.012	-0.057	-0.070
15	6	3	3	-0.021	-0.085	-0.069
24 Total Items, 12 Items Taken						
16	6	6	2	-0.103	-0.008	-0.029
17	6	6	4	-0.059	0.006	-0.000
18	6	8	2	-0.150	0.012	-0.025
19	6	8	4	-0.117	-0.017	-0.059
20	12	6	6	-0.062	0.030	0.000

Note: Standard errors are typically less than 0.025

Table A.18 48 Total Items: Scale Means (100*Standard Error) for Two Equating Methods (Mathematics Simulation)

Cell	Number of: Common Items	Type I Items	Common Type I Items	Wave 1	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Common Item Equating
12 Items Taken							
1	6	6	1	0.024 (0.205)	0.021 (0.220)	0.011 (0.361)	-0.129 (0.346)
2	6	6	3	0.021 (0.141)	0.021 (0.254)	0.017 (0.296)	-0.124 (0.258)
3	6		7	0.023 (0.160)	0.020 (0.246)	0.012 (0.373)	-0.121 (0.317)
4	6	8	3	0.021 (0.134)	0.015 (0.339)	0.016 (0.262)	-0.122 (0.204)
5	12	6	6	0.023 (0.191)	0.023 (0.305)	0.022 (0.200)	-0.135 (0.359)
24 Items Taken							
6	12	12	2	0.020 (0.168)	0.016 (0.152)	0.004 (0.290)	-0.128 (0.313)
7	12	12	6	0.019 (0.120)	0.019 (0.170)	0.009 (0.363)	-0.123 (0.181)
8	12	14	2	0.019 (0.164)	0.019 (0.244)	0.006 (0.330)	-0.130 (0.210)
9	12	16	6	0.018 (0.084)	0.017 (0.294)	0.006 (0.268)	-0.126 (0.241)
10	24	12	12	0.020 (0.091)	0.022 (0.300)	0.013 (0.336)	-0.137 (0.266)

Table A.19 24 Total Items: Scale Means (100*Standard Error) for Two Equating Methods (Mathematics Simulation)

Cell	Number of: Common Common Items	Type I Items	Common Type I Items	Wave 1	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Common Item Equating
6 Items Taken							
11	3	3	1	0.016 (0.156)	0.017 (0.417)	0.016 (0.218)	-0.136 (0.272)
12	3	3	2	0.015 (0.278)	0.016 (0.304)	0.018 (0.302)	-0.125 (0.378)
13	3	4	1	0.019 (0.213)	0.021 (0.270)	0.017 (0.166)	-0.130 (0.316)
14	3	4	2	0.021 0.145)	0.017 (0.259)	0.016 (0.192)	-0.128 (0.225)
15	6	3	3	0.015 (0.234)	0.018 (0.354)	0.022 (0.186)	-0.131 (0.187)
12 Items Taken							
16	6	6	2	0.023 (0.223)	0.040 (2.220)	0.014 (0.257)	-0.119 (0.383)
17	6	6	4	0.022 (0.213)	0.053 (1.832)	0.018 (0.254)	0.107 (0.253)
18	6	8	2	0.025 0.167)	0.046 (1.686)	0.014 (0.488)	-0.126 (0.424)
19	6	8	4	0.024 (0.117)	0.073 (2.253)	0.020 (0.265)	-0.108 (0.199)
20	12	6	6	0.020 (0.201)	0.042 (1.359)	0.024 (0.300)	-0.133 (0.215)

Table A.20 48 Total Items: Scale Variances (100*Standard Error) for Two Equating Methods (Mathematics Simulation)

Cell	Number of: Common Common Items	Type I Type I Items	Common Type I Items	Wave 1	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Common Item Equating
12 Items Taken							
1	6	6	1	1.140 (0.356)	1.143 (0.567)	1.197 (0.457)	1.188 (0.516)
2	6	6	3	1.151 (0.305)	1.141 (0.450)	1.194 (0.281)	1.193 (0.572)
3	6	7	1	1.142 (0.240)	1.139 (0.547)	1.195 (0.400)	1.185 (0.478)
4	6	8	3	1.135 (0.297)	1.119 (0.578)	1.183 (0.301)	1.175 (0.272)
5	12	6	6	1.154 (0.409)	1.137 (0.644)	1.199 (0.478)	1.187 (0.684)
24 Items Taken							
6	12	12	2	1.147 (0.314)	1.143 (0.421)	1.179 (0.437)	1.166 (0.415)
7	12	12	6	1.147 (0.292)	1.146 (0.571)	1.177 (0.434)	1.173 (0.432)
8	12	14	2	1.143 (0.347)	1.142 (0.441)	1.174 (0.763)	1.162 (0.465)
9	12	16	6	1.141 (0.279)	1.135 (0.445)	1.161 (0.399)	1.171 (0.304)
10	24	12	12	1.143 (0.228)	1.135 (0.280)	1.182 (0.620)	1.165 (0.426)

Table A.21 24 Total Items: Scale Variences (100*Standard Error) for Two Equating Methods (Mathematics Simulation)

Cell	Number of: Common Common Items	Type I Items	Common Type I Items	Wave 1	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Common Item Equating
6 Items Taken							
11	3	3	1	1.100 (0.279)	1.093 (0.301)	1.119 (0.437)	1.116 (0.319)
12	3	3	2	1.087 (0.330)	1.104 (0.643)	1.116 (0.629)	1.120 (0.504)
13	3	4	1	1.106 (0.346)	1.080 (1.047)	1.127 (0.275)	1.109 (0.520)
14	3	4	2	1.096 (0.615)	1.098 (0.400)	1.126 (0.424)	1.115 (0.444)
15	6	3	3	1.095 (0.352)	1.097 (0.263)	1.067 (0.386)	1.061 (0.313)
12 Items Taken							
16	6	6	2	1.146 (0.201)	0.990 (3.586)	1.207 (0.348)	1.188 (0.326)
17	6	6	4	1.143 (0.183)	1.020 (5.973)	1.209 (0.422)	1.199 (0.366)
18	6	8	2	1.138 (0.328)	0.909 (4.178)	1.195 (0.457)	1.195 (0.395)
19	6	8	4	1.145 (0.282)	0.929 (7.167)	1.194 (0.726)	1.189 (0.309)
20	12	6	6	1.143 (0.301)	1.024 (3.513)	1.195 (0.466)	1.190 (0.434)

Table A.22 48 Total Items Scale Skewness (100*Standard Error) for Two Equating Methods (Mathematics Simulation)

Cell	Number of: Common Common Items	Number of: Type I Type I Items	Number of: Common Type I Items	Wave 1	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Common Item Equating
12 Items Taken							
1	6	6	1	0.066 (0.368)	0.064 (0.732)	0.090 (0.516)	0.097 (0.590)
2	6	6	3	0.072 (0.911)	0.082 (0.467)	0.083 (0.689)	0.095 (0.729)
3	6	7	1	0.071 (0.621)	0.072 (0.308)	0.086 (0.612)	0.099 (0.866)
4	6	8	3	0.068 (0.437)	0.070 (0.686)	0.084 (0.665)	0.093 (0.374)
5	12	6	6	0.058 (0.539)	0.066 (0.781)	0.094 (0.809)	0.088 (0.716)
24 Items Taken							
6	12	12	2	0.023 (0.340)	0.037 (0.781)	0.068 (0.947)	0.065 (0.553)
7	12	12	6	0.030 (0.707)	0.036 (0.609)	0.057 (0.541)	0.070 (0.634)
8	12	14	2	0.037 (0.909)	0.049 (0.522)	0.078 (0.395)	0.072 (0.968)
9	12	16	6	0.038 (0.567)	0.036 (1.042)	0.068 (0.559)	0.068 (0.609)
10	24	12	12	0.026 (0.721)	0.033 (0.725)	0.036 (0.402)	0.041 (0.708)

Table A.23 24 Total Items: Scale Skewness (100*Standard Error) for Two Equating Methods (Mathematics Simulation)

Cell	Number of Common Items	Number of Type I Items	Number of Common Type I Items	Wave 1	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Common Item Equating
6 Items Taken							
11	3	3	1	0.045 (0.827)	0.060 (0.826)	0.042 (0.759)	0.042 (0.628)
12	3	3	2	0.039 (0.784)	0.037 (0.522)	0.035 (0.725)	0.044 (0.662)
13	3	4	1	0.065 (0.510)	0.041 (0.684)	0.043 (0.524)	0.035 (0.463)
14	3	4	2	0.049 (0.603)	0.047 (0.686)	0.041 (0.633)	0.046 (0.516)
15	6	3	3	0.038 (0.704)	0.052 (0.430)	0.040 (0.324)	0.025 (0.690)
12 Items Taken							
16	6	6	2	0.051 (0.441)	-0.008 (1.319)	0.069 (0.330)	0.073 (0.405)
17	6	6	4	0.060 (0.551)	-0.010 (3.538)	0.078 (0.759)	0.069 (0.986)
18	6	8	2	0.055 (0.334)	-0.024 (2.399)	0.070 (0.440)	0.094 (0.361)
19	6	8	4	0.066 (0.375)	-0.075 (4.599)	0.071 (0.814)	0.085 (0.777)
20	12	6	6	0.050 (0.770)	-0.018 (1.698)	0.072 (0.822)	0.082 (0.640)

Table A.24 48 Total Items: Scale Kurtosis (100*Standard Error) for Two Equating Methods (Mathematics Simulation)

Cell	Number of: Common Items	Type I Items	Common Type I Items	Wave 1	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Common Item Equating
12 Items Taken							
1	6	6	1	-0.148 (1.193)	-0.159 (0.735)	-0.152 (0.938)	-0.141 (1.249)
2	6	6	3	-0.167 (1.087)	-0.150 (1.763)	-0.168 (1.135)	-0.132 (1.459)
3	6	7	1	-0.175 (1.082)	-0.162 (0.765)	-0.163 (0.859)	-0.153 (1.021)
4	6	8	3	-0.171 (1.210)	-0.153 (1.362)	-0.132 (1.687)	-0.150 (1.099)
5	12	6	6	-0.169 (1.088)	-0.148 (1.115)	-0.145 (1.158)	-0.127 (1.053)
24 Items Taken							
6	12	12	2	-0.241 (1.138)	-0.263 (1.524)	-0.245 (0.840)	-0.218 (1.262)
7	12	12	6	-0.238 (1.513)	-0.262 (1.253)	-0.262 (1.086)	-0.229 (1.021)
8	12	14	2	-0.255 (0.539)	-0.246 (1.448)	-0.249 (1.491)	-0.232 (1.383)
9	12	16	6	-0.245 (1.079)	-0.219 (1.488)	-0.230 (1.094)	-0.211 (0.940)
10	24	12	12	-0.260 (0.687)	-0.243 (0.993)	-0.276 (0.721)	-0.260 (1.142)

Table A.25 24 Total Items: Scale Kurtosis (100*Standard Error) for Two Equating Methods (Mathematics Simulation)

Cell	Number of: Common Items	Type I Items	Common Type I Items	Wave 1	No Population Change: Invariance Equating	No Population Change: Common Item Equating	-0.15 Population Change: Common Item Equating
6 Items Taken							
11	3	3	1	-0.105 (0.962)	-0.090 (0.733)	-0.078 (0.954)	-0.035 (0.855)
12	3	3	2	-0.092 (1.071)	-0.111 (1.071)	-0.046 (0.833)	-0.045 (1.351)
13	3	4	1	-0.096 (1.526)	-0.095 (1.295)	-0.077 (1.142)	-0.051 (0.943)
14	3	4	2	-0.116 (1.407)	-0.104 (0.492)	-0.059 (0.714)	-0.046 (1.110)
15	6	3	3	-0.100 (1.270)	-0.079 (1.406)	-0.015 (1.117)	-0.031 (0.969)
12 Items Taken							
16	6	6	2	-0.181 (1.030)	-0.078 (2.592)	-0.173 (1.237)	-0.152 (0.893)
17	6	6	4	-0.174 (0.959)	-0.115 (2.834)	-0.180 (1.053)	-0.174 (0.816)
18	6	8	2	-0.174 (1.263)	-0.024 (4.074)	-0.186 (1.026)	-0.149 (1.427)
19	6	8	4	-0.208 (0.707)	-0.091 (3.613)	-0.192 (0.820)	-0.149 (1.239)
20	12	6	6	-0.156 (0.999)	-0.094 (3.229)	-0.186 (0.687)	-0.156 (0.945)

Table A.26 Average Scale Linking Bias Using Multiple Group IRT

	Number of:			No Population Change		-0.15 Population Change	
Cell	Common Items	Type I Items	Common Type I Items	Reading	Math	Reading	Math
48 Total Items, 12 Items Taken							
1	6	6	1	-0.003	-0.001	-0.003	-0.002
2	6	6	3	0.003	-0.003	0.000	-0.007
3	6	7	1	0.001	0.005	-0.014	0.002
4	6	8	3	0.001	-0.008	0.001	-0.002
5	12	6	6	-0.005	-0.004	-0.003	0.006
48 Total Items, 24 Items Taken							
6	12	12	2	-0.007	-0.002	0.001	-0.005
7	12	12	6	-0.008	0.002	-0.004	0.000
8	12	14	2	-0.011	0.001	-0.003	0.002
9	12	16	6	0.005	-0.004	-0.001	-0.010
10	24	12	12	-0.001	0.001	0.000	0.001
24 Total Items, 6 Items Taken							
11	3	3	1	0.001	-0.004	-0.006	-0.003
12	3	3	2	0.000	0.003	0.002	0.005
13	3	4	1	-0.001	-0.009	-0.001	-0.002
14	3	4	2	-0.001	0.004	-0.011	0.004
15	6	3	3	0.002	-0.000	0.000	
24 Total Items, 12 Items Taken							
16	6	6	2	0.004	0.001	0.003	0.001
17	6	6	4	0.006	-0.004	0.005	-0.007
18	6	8	2	-0.006	-0.005	-0.008	-0.002
19	6	8	4	-0.003	0.012	-0.004	0.005
20	12	6	6	-0.005	0.004	0.000	-0.001

Note: Standard errors are typically less than 0.003.

Table A.27. Ratio of Wave 2 to Wave 1 Variance Using Multiple Group IRT

Cell	Number of: Common Items	Type I Items	Common Type I Items	No Population Change: Reading	No Population Change: Math	-0.15 Population Change: Reading	-0.15 Population Change: Math
48 Total Items, 12 Items Taken							
1	6	6	1	1.002	0.988	0.991	0.983
2	6	6	3	0.996	0.991	0.998	0.985
3	6	7	1	0.993	0.988	1.001	0.995
4	6	8	3	0.993	1.000	0.996	0.994
5	12	6	6	1.001	1.001	1.002	0.992
48 Total Items, 24 Items Taken							
6	12	12	2	0.989	0.985	0.995	0.977
7	12	12	6	1.009	1.000	1.003	0.979
8	12	14	2	0.992	0.981	0.990	0.983
9	12	16	6	0.981	0.991	0.990	0.982
10	24	12	12	0.995	1.010	0.999	0.996
24 Total Items, 6 Items Taken							
11	3	3	1	1.008	1.005	0.994	0.992
12	3	3	2	1.008	0.997	1.003	0.988
13	3	4	1	1.001	0.990	0.998	0.983
14	3	4	2	0.993	1.002	0.995	0.994
15	6	3	3	0.994	1.005	0.997	
24 Total Items, 12 Items Taken							
16	6	6	2	1.009	1.010	1.012	0.986
17	6	6	4	0.998	0.993	1.007	0.995
18	6	8	2	0.990	0.990	0.991	0.977
19	6	8	4	1.002	0.999	0.995	0.990
20	12	6	6	1.004	1.006	0.997	0.988

Table A.28 Difference Between Wave 2 and Wave 1 Skewness Using Multiple Group IRT

	Number of:			No Population Change		-0.15 Population Change	
Cell	Common Items	Type I Items	Common Type I Items	Reading	Math	Reading	Math
48 Total Items, 12 Items Taken							
1	6	6	1	0.009	-0.021	-0.001	-0.008
2	6	6	3	0.002	0.003	0.001	0.004
3	6	7	1	-0.010	0.001	-0.008	-0.009
4	6	8	3	0.002	-0.028	-0.005	-0.011
5	12	6	6	0.003	0.001	-0.008	0.004
48 Total Items, 24 Items Taken							
6	12	12	2	-0.003	-0.045	0.000	-0.023
7	12	12	6	0.003	-0.026	-0.010	-0.007
8	12	14	2	-0.027	-0.068	-0.032	-0.047
9	12	16	6	-0.001	-0.010	-0.002	-0.023
10	24	12	12	-0.009	-0.027	-0.021	-0.011
24 Total Items, 6 Items Taken							
11	3	3	1	0.010	0.018	-0.004	0.011
12	3	3	2	-0.010	-0.014	-0.018	0.010
13	3	4	1	-0.009	0.011	-0.003	0.003
14	3	4	2	0.002	0.006	-0.009	-0.002
15	6	3	3	0.011	-0.003	-0.003	
24 Total Items, 12 Items Taken							
16	6	6	2	0.004	-0.004	0.000	-0.000
17	6	6	4	-0.003	0.010	-0.020	0.006
18	6	8	2	-0.015	0.010	0.004	-0.011
19	6	8	4	-0.003	0.015	-0.020	0.016
20	12	6	6	-0.017	-0.003	-0.005	-0.000

Note: Standard errors are typically less than 0.008.

Table A.29 Difference Between Wave 2 and Wave 1 Kurtosis Using Multiple Group IRT

	Number of:			No Population Change		-0.15 Population Change	
Cell	Common Items	Type I Items	Common Type I Items	Reading	Math	Reading	Math
48 Total Items, 12 Items Taken							
1	6	6	1	0.007	-0.030	0.016	-0.016
2	6	6	3	-0.000	0.006	0.002	-0.023
3	6	7	1	0.014	-0.039	-0.020	-0.043
4	6	8	3	-0.024	-0.020	0.005	-0.011
5	12	6	6	-0.018	-0.006	-0.020	-0.005
48 Total Items, 24 Items Taken							
6	12	12	2	0.031	-0.012	0.013	-0.039
7	12	12	6	0.008	0.003	-0.011	0.005
8	12	14	2	0.044	-0.038	0.018	-0.027
9	12	16	6	-0.035	0.003	0.001	-0.022
10	24	12	12	0.004	-0.005	0.019	-0.019
24 Total Items, 6 Items Taken							
11	3	3	1	-0.006	0.012	0.015	-0.013
12	3	3	2	0.002	-0.019	-0.001	-0.008
13	3	4	1	0.010	0.010	-0.012	-0.009
14	3	4	2	-0.005	-0.012	-0.007	-0.025
15	6	3	3	-0.000	-0.009	-0.047	
24 Total Items, 12 Items Taken							
16	6	6	2	-0.003	-0.004	-0.011	-0.029
17	6	6	4	0.002	0.016	0.006	0.003
18	6	8	2	-0.019	-0.019	0.021	-0.035
19	6	8	4	0.000	-0.007	0.026	-0.001
20	12	6	6	0.039	-0.012	0.020	-0.025

Note: Standard errors are typically less than 0.025.

Section B: Figures

Figure B.1a. Histogram of the Generating Populations of Abilities: Reading

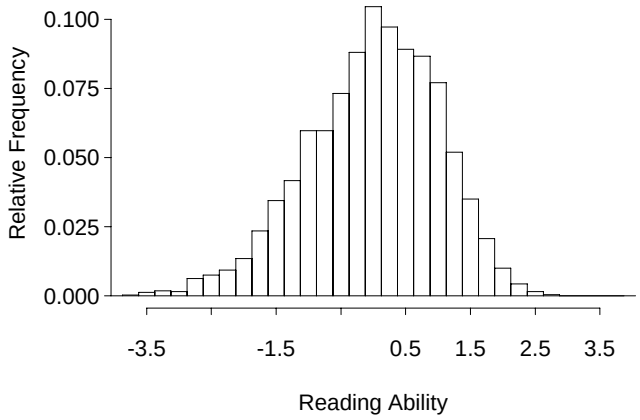


Figure B.1b. Histogram of the Generating Populations of Abilities: Mathematics

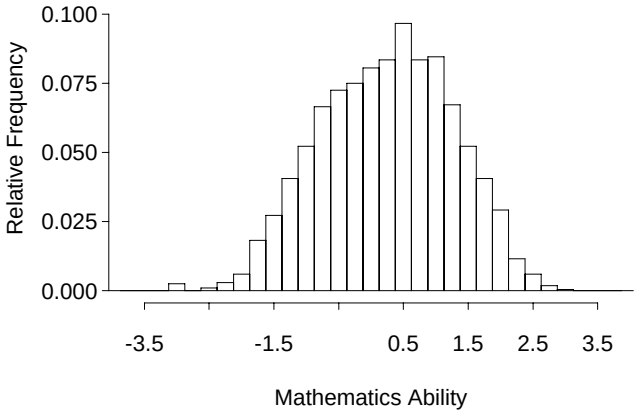


Figure B.2. Quantiles of Scale Distributions for Cell 1 (Reading Simulation)

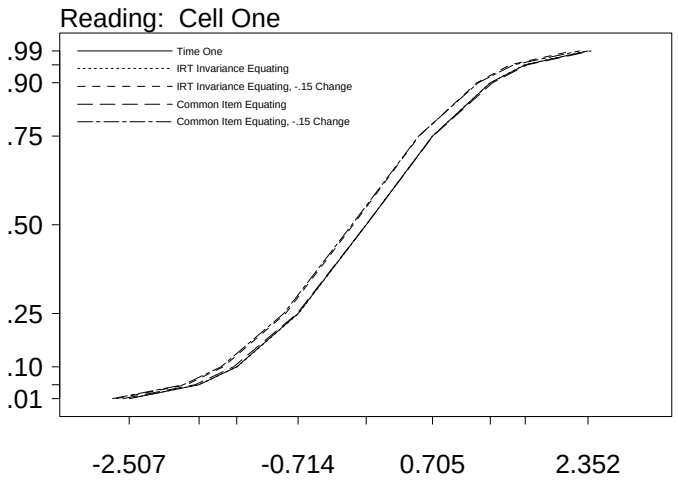


Figure B.3. Quantiles of Scale Distributions for Cell 2 (Reading Simulation)

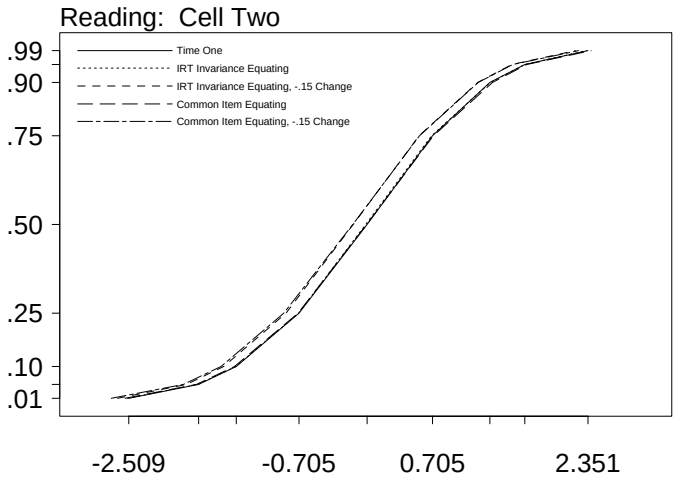


Figure B.4. Quantiles of Scale Distributions for Cell 3 (Reading Simulation)

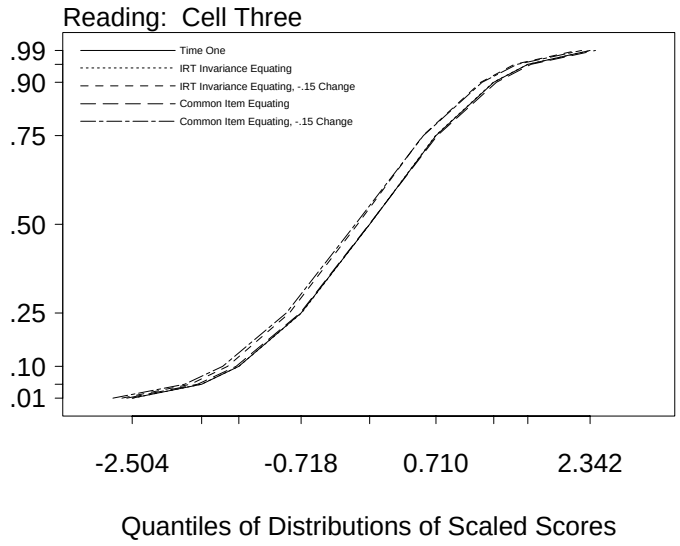


Figure B.5. Quantiles of Scale Distributions for Cell 4 (Reading Simulation)

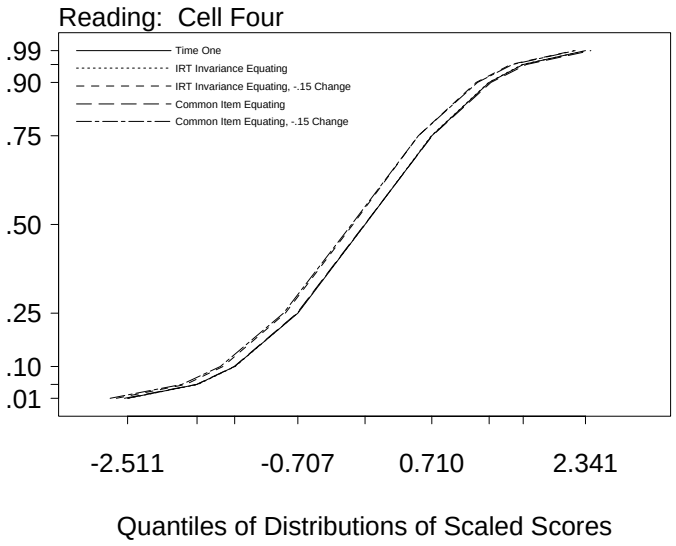


Figure B.6. Quantiles of Scale Distributions for Cell 5 (Reading Simulation)

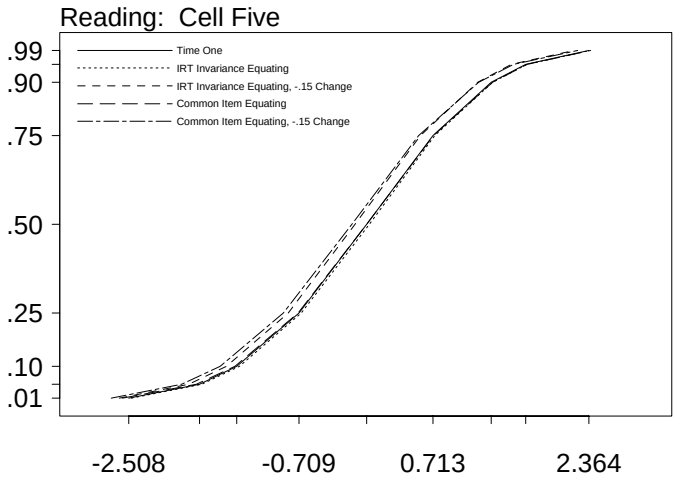


Figure B.7. Quantiles of Scale Distributions for Cell 6 (Reading Simulation)

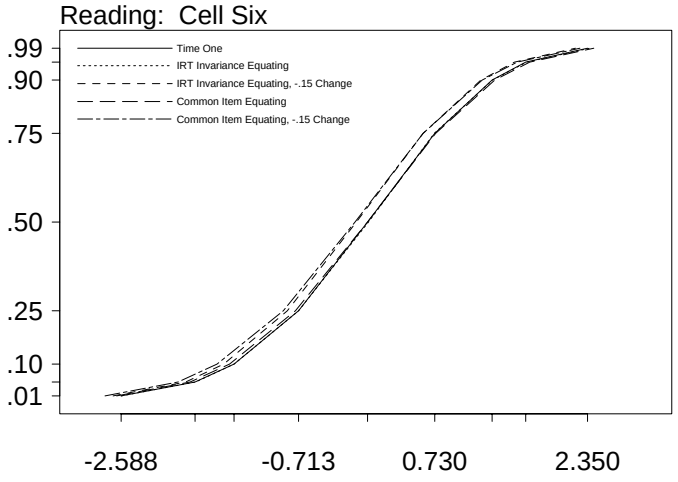


Figure B.8. Quantiles of Scale Distributions for Cell 7 (Reading Simulation)

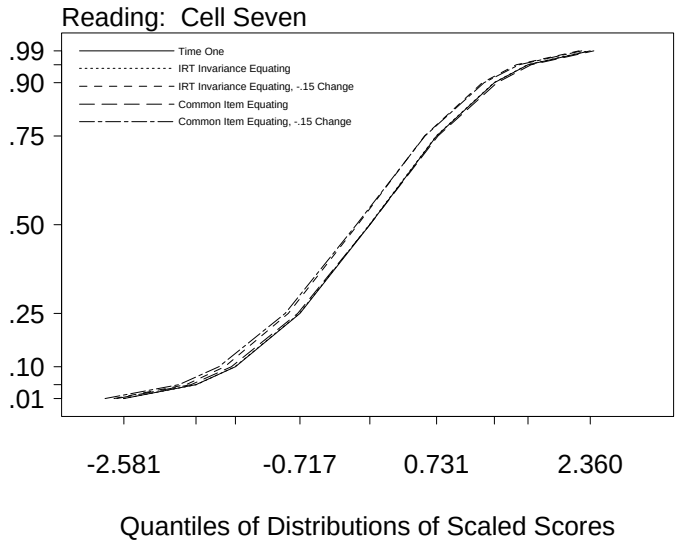


Figure B.9. Quantiles of Scale Distributions for Cell 8 (Reading Simulation)

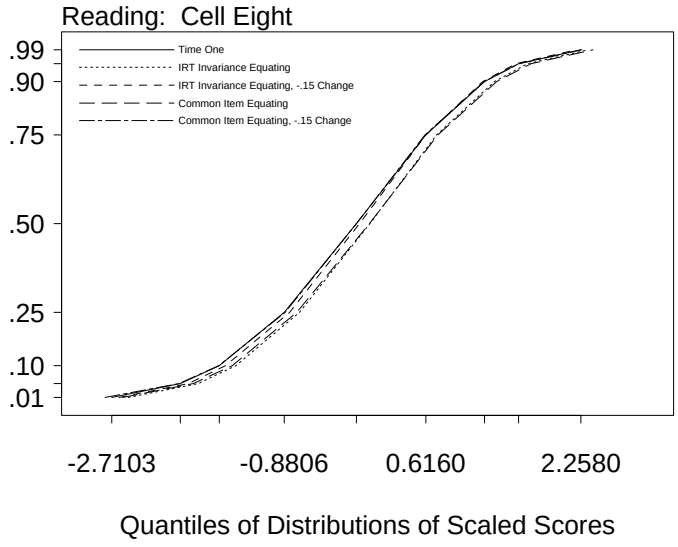


Figure B.10. Quantiles of Scale Distributions for Cell 9 (Reading Simulation)

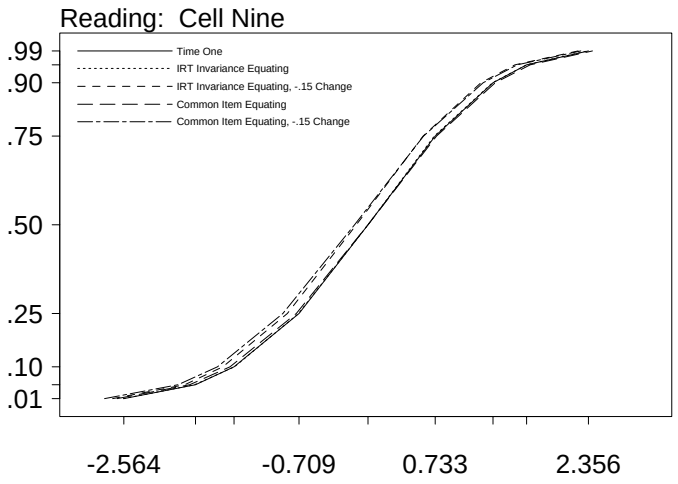


Figure B.11. Quantiles of Scale Distributions for Cell 10 (Reading Simulation)

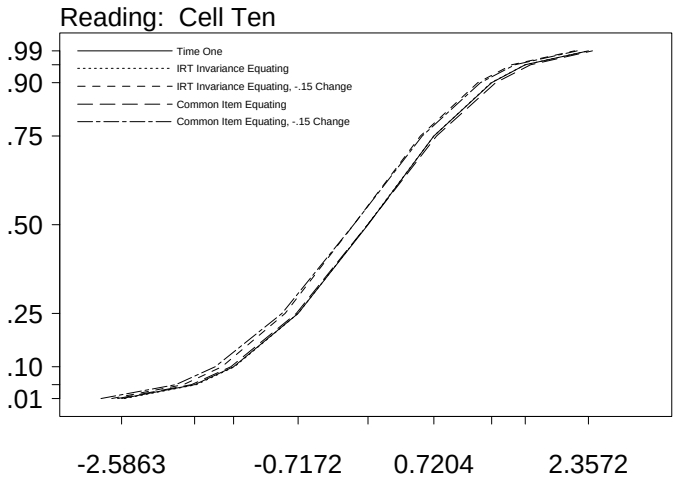


Figure B.12. Quantiles of Scale Distributions for Cell 11 (Reading Simulation)

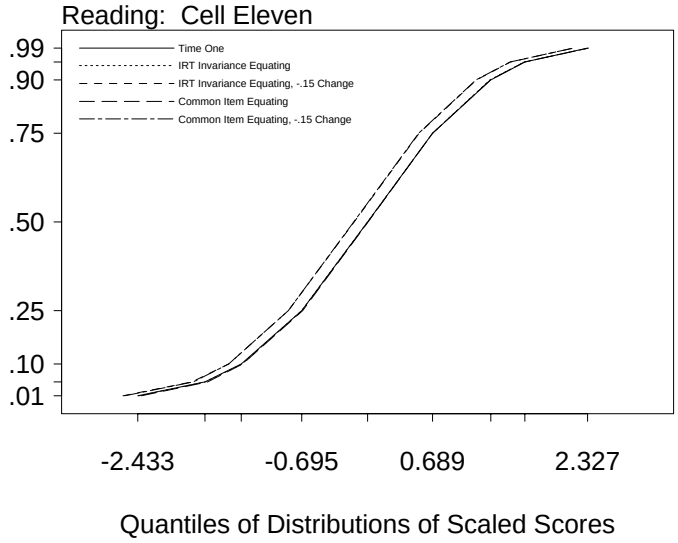


Figure B.13. Quantiles of Distribution for Cell 12 (Reading Simulation)

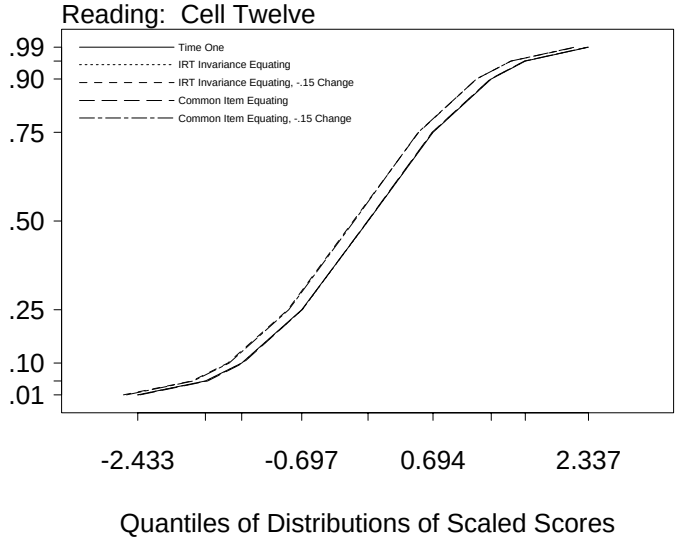


Figure B.14. Quantiles of Scale Distributions for Cell 13 (Reading Simulation)

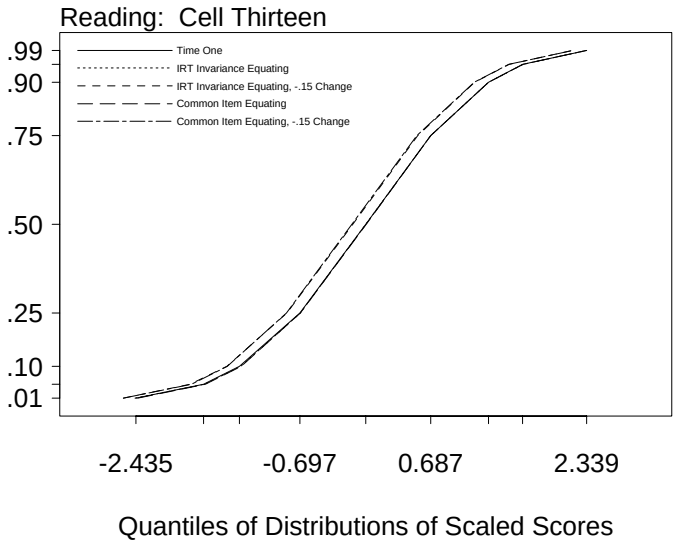


Figure B.15. Quantiles of Scale Distributions for Cell 14 (Reading Simulation)

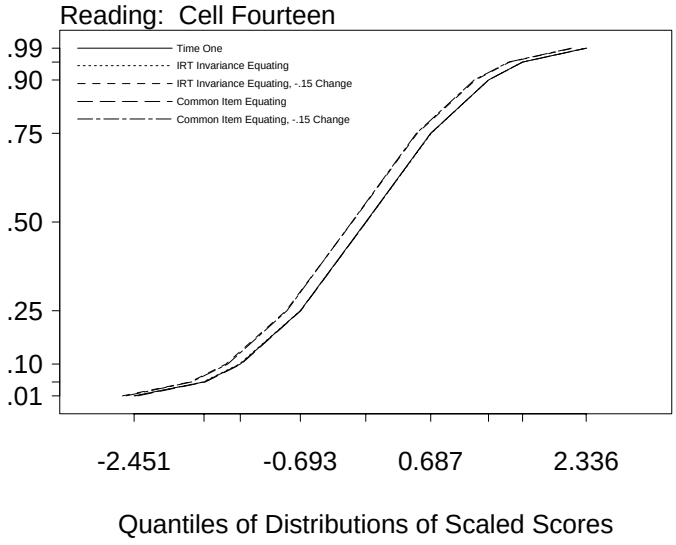


Figure B.16. Quantiles of Scale Distributions for Cell 15 (Reading Simulation)

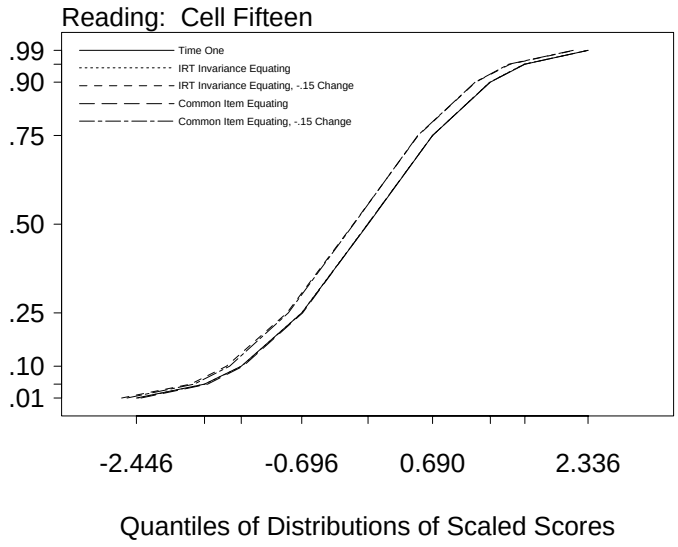


Figure B.17. Quantiles of Scale Distributions for Cell 16 (Reading Simulation)

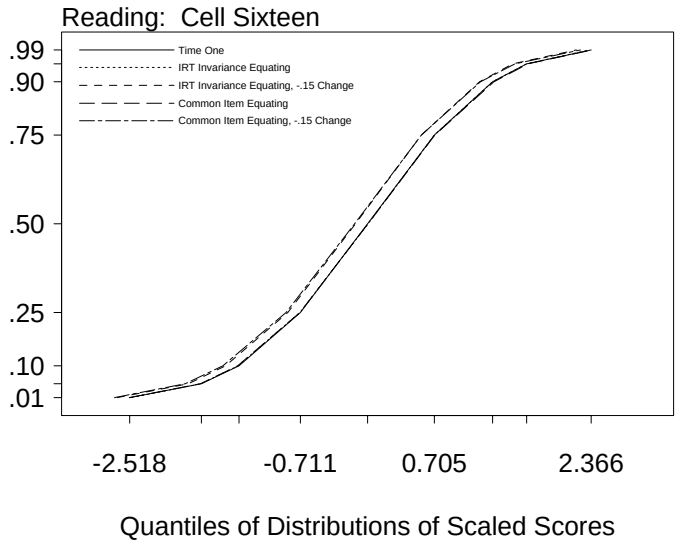


Figure B.18. Quantiles of Scale Distributions for Cell 17 (Reading Simulation)

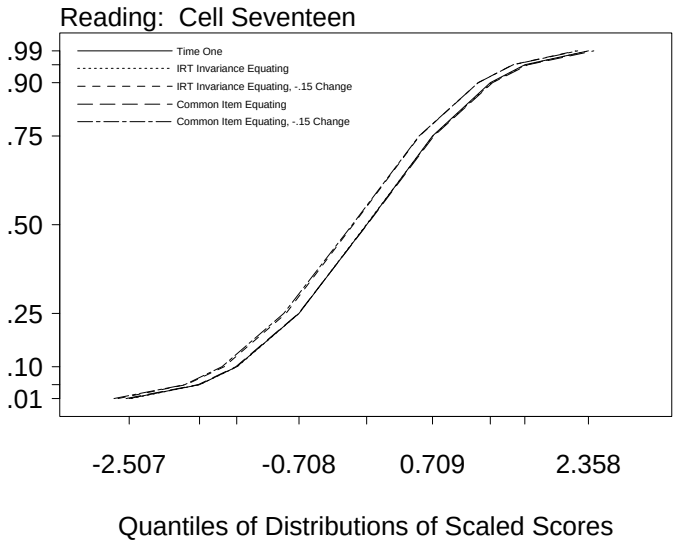


Figure B.19. Quantiles of Scale Distributions for Cell 18 (Reading Simulation)

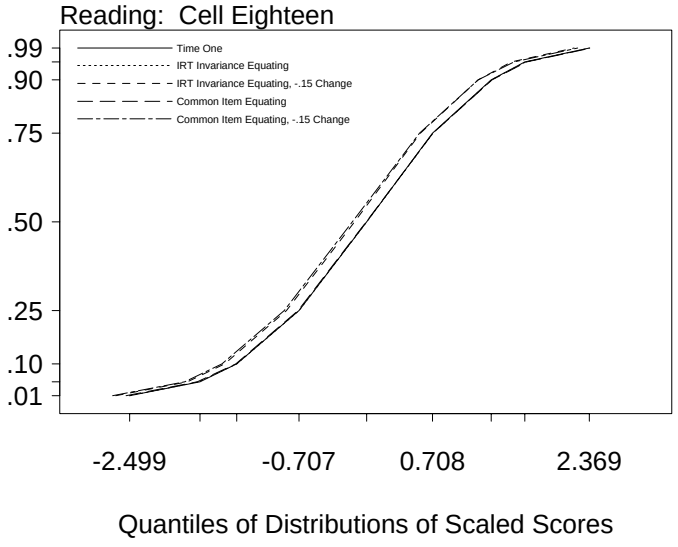


Figure B.20. Quantiles of Scale Distributions for Cell 19 (Reading Simulation)

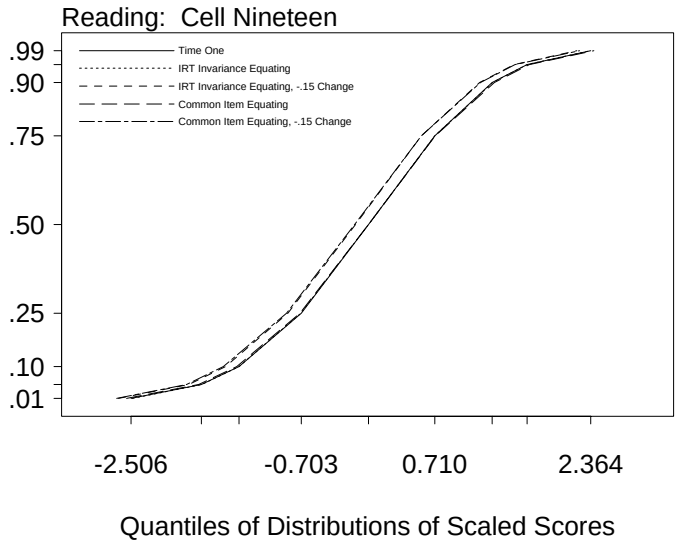


Figure B.21. Quantiles of Scale Distributions for Cell 20 (Reading Simulation)

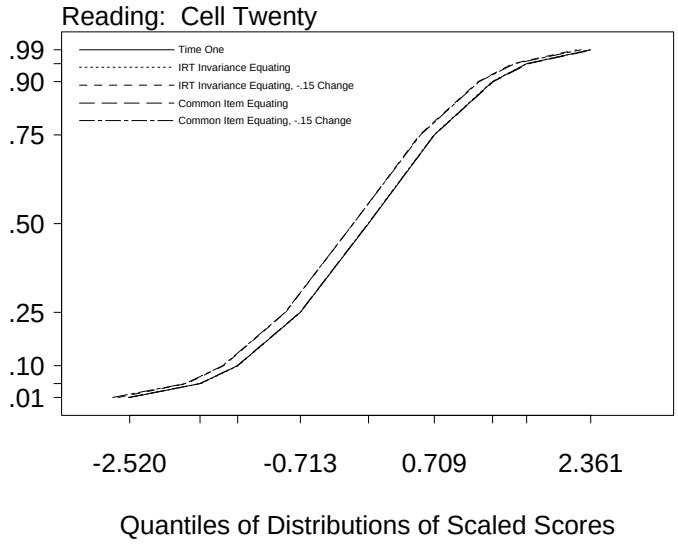


Figure B.22. Quantiles of Scale Distributions for Cell Eight (Reading Simulation), with Increased Multidimensionality $\lambda = 0.7$

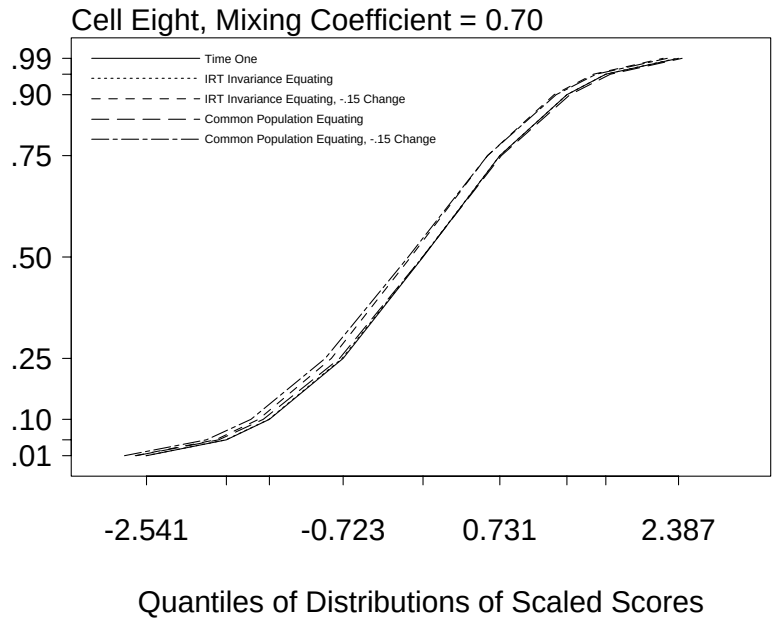


Figure B.23. Quantiles of Scale Distributions for Cell 1 (Mathematics Simulation)

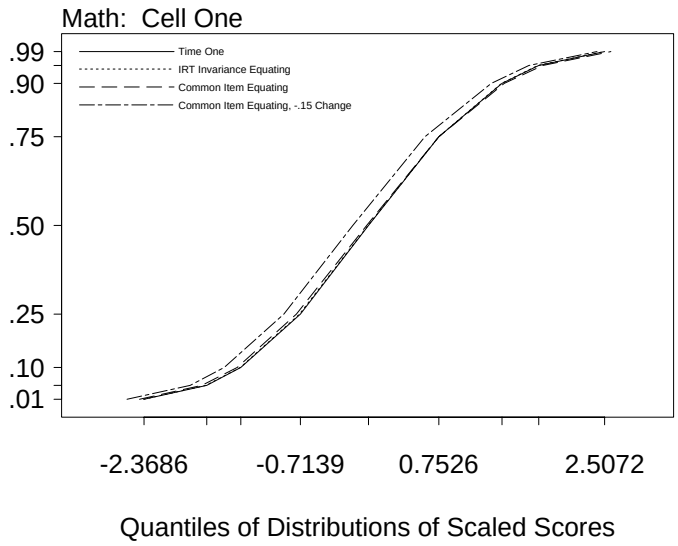


Figure B.24. Quantiles of Scale Distributions for Cell 2 (Mathematics Simulation)

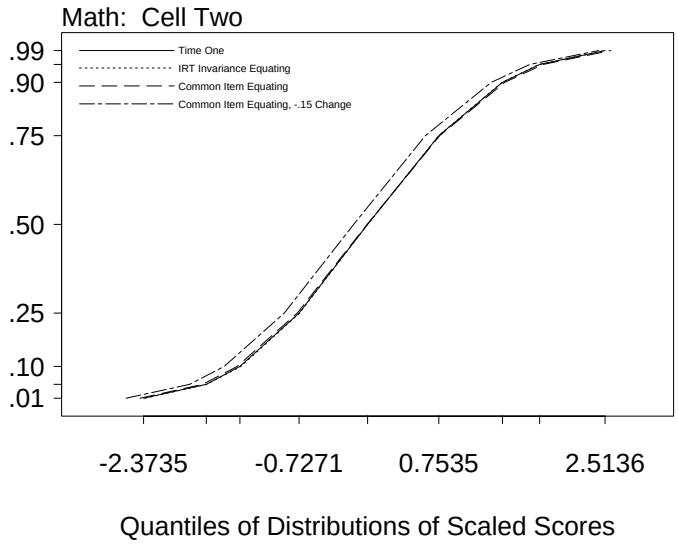


Figure B.25. Quantiles of Scale Distributions for Cell 3 (Mathematics Simulation)

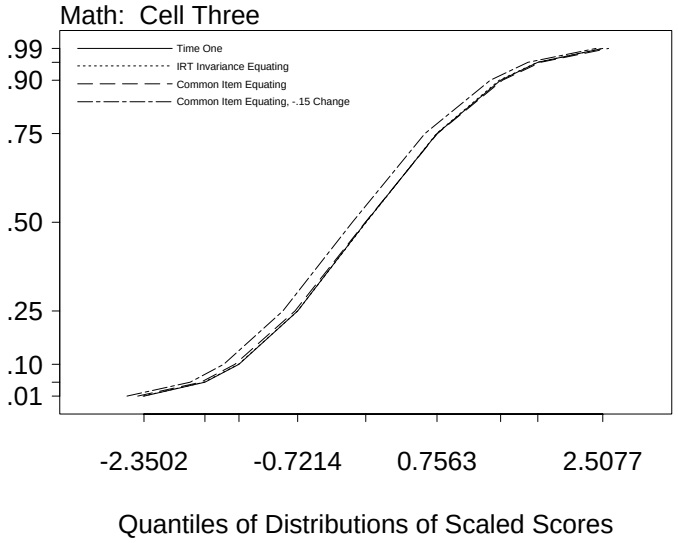


Figure B.26. Quantiles of Scale Distributions for Cell 4 (Mathematics Simulation)

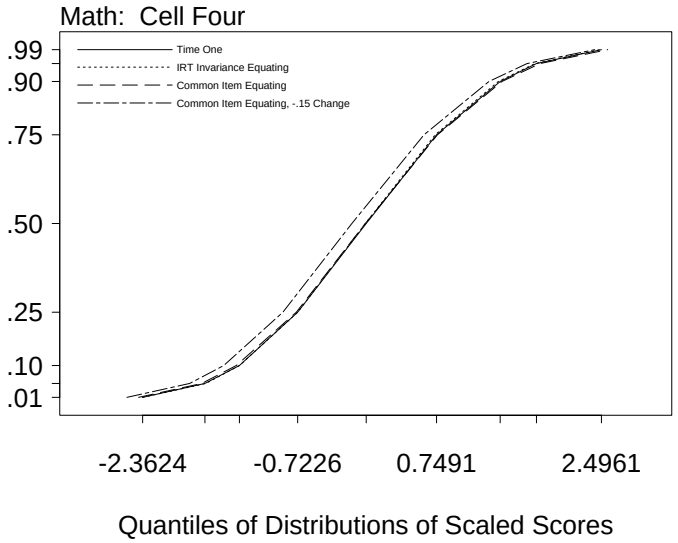


Figure B.27. Quantiles of Scale Distributions for Cell 5 (Mathematics Simulation)

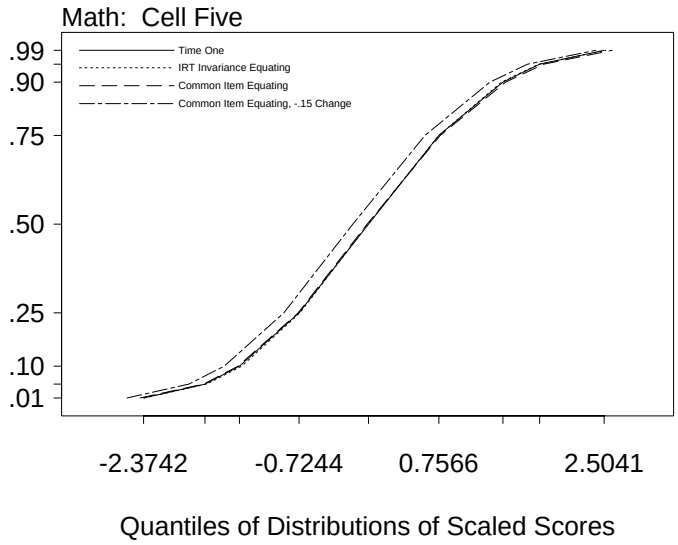


Figure B.28. Quantiles of Scale Distributions for Cell 6 (Mathematics Simulation)

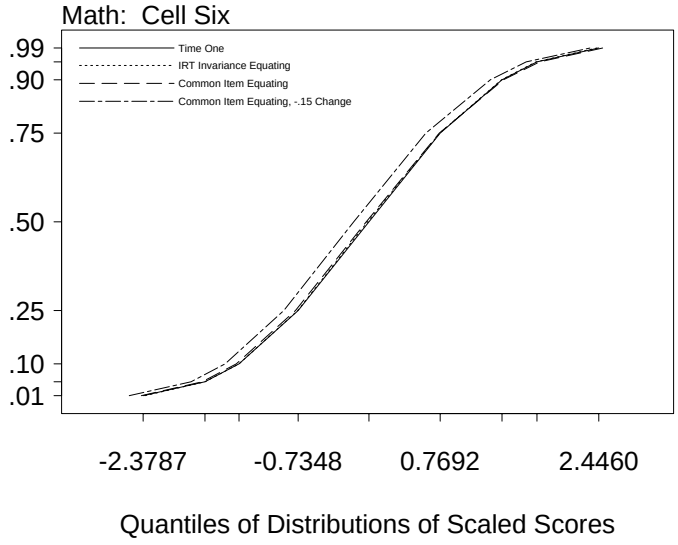


Figure B.29. Quantiles of Scale Distributions for Cell 7 (Mathematics Simulation)

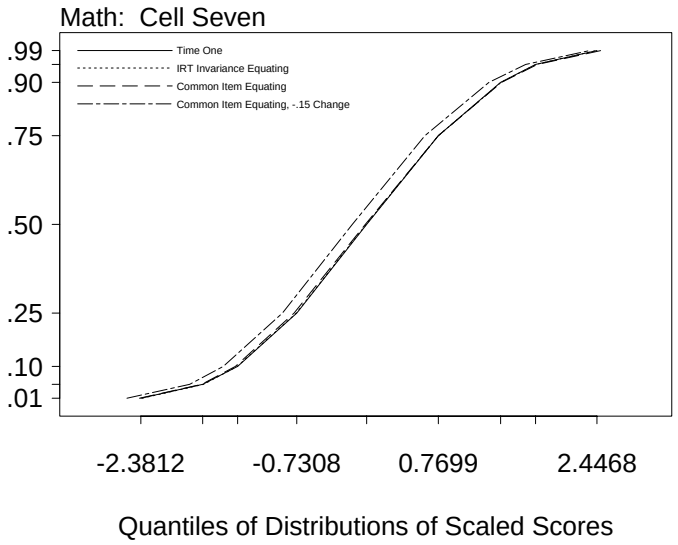


Figure B.30. Quantiles of Scale Distributions for Cell 8 (Mathematics Simulation)

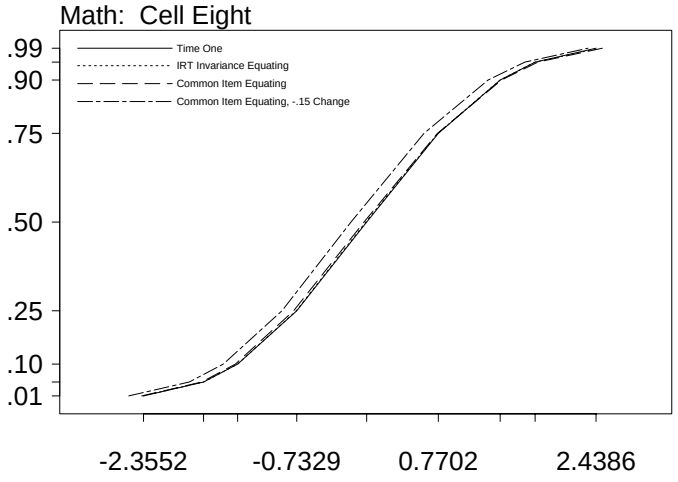


Figure B.31. Quantiles of Scale Distributions for Cell 9 (Mathematics Simulation)

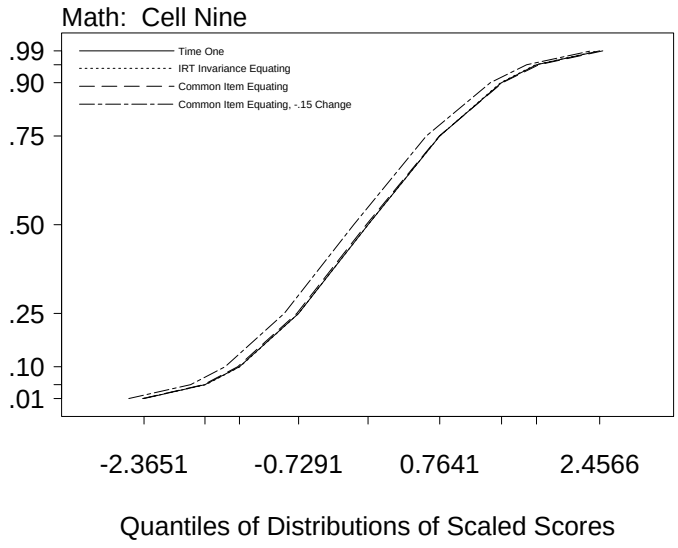


Figure B.32. Quantiles of Scale Distributions for Cell 10 (Mathematics Simulations)

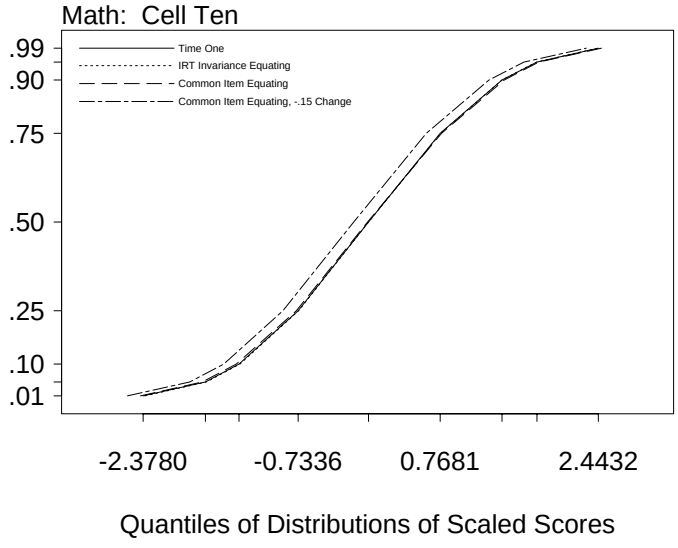


Figure B.33. Quantiles of Scale Distributions for Cell 11 (Mathematics Simulation)

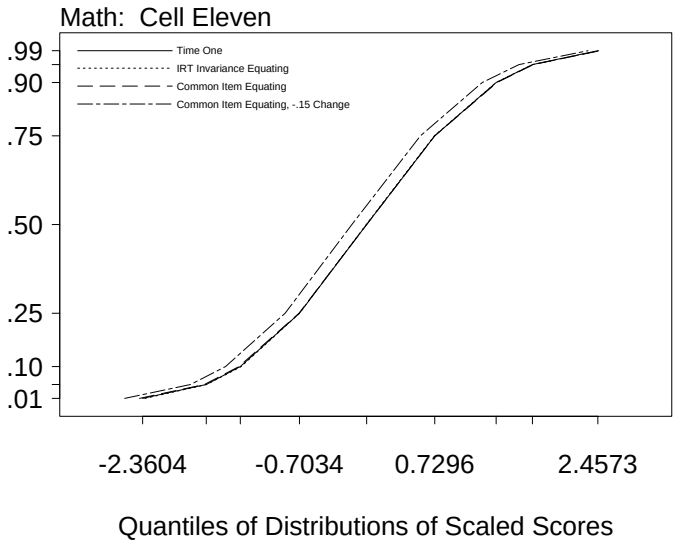


Figure B.34. Quantiles of Scale Distributions for Cell 12 (Mathematics Simulation)

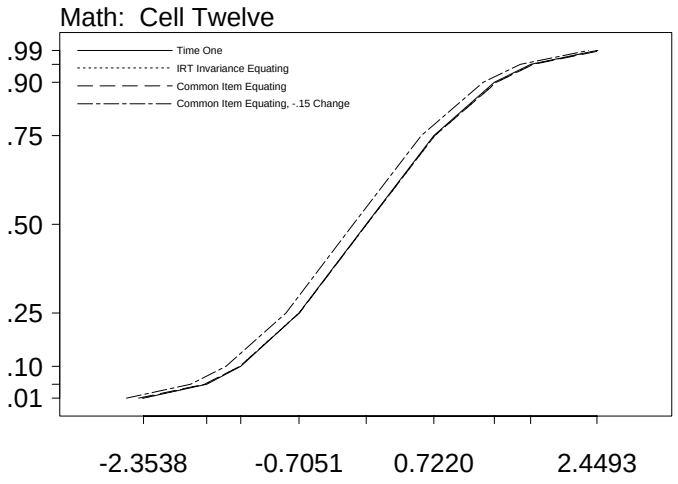


Figure B.35. Quantiles of Scale Distributions of Cell 13 (Mathematics Simulation)

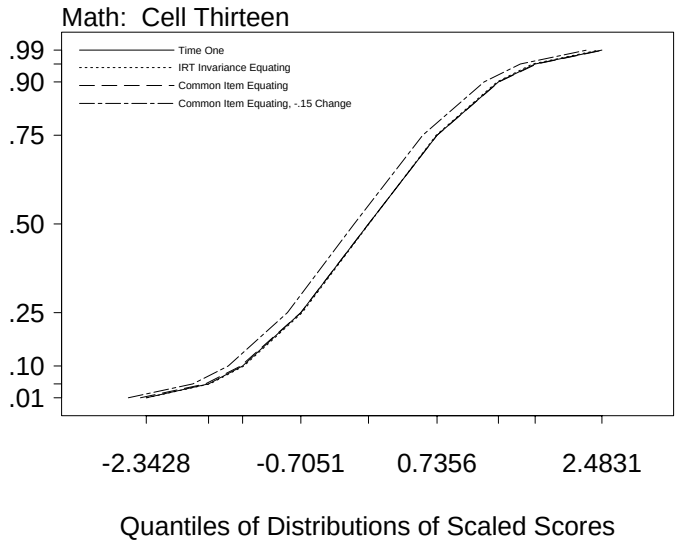


Figure B.36. Quantiles of Scale Distributions of Cell 14 (Mathematics Simulation)

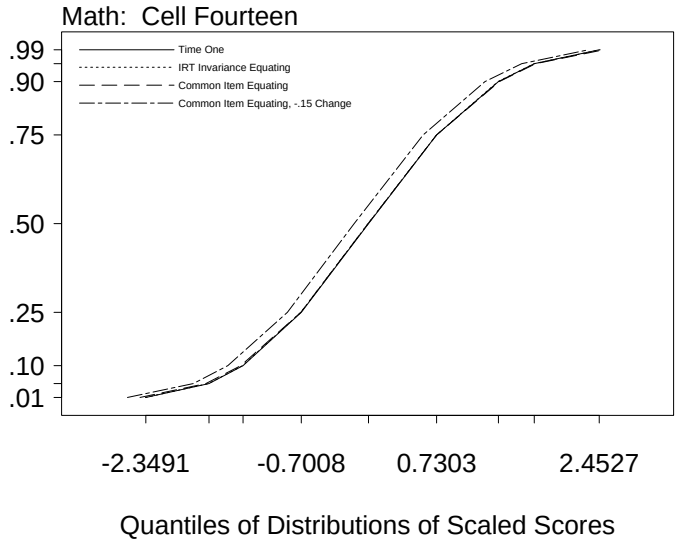


Figure B.37. Quantiles of Scale Distributions for Cell 15 (Mathematics Simulation)

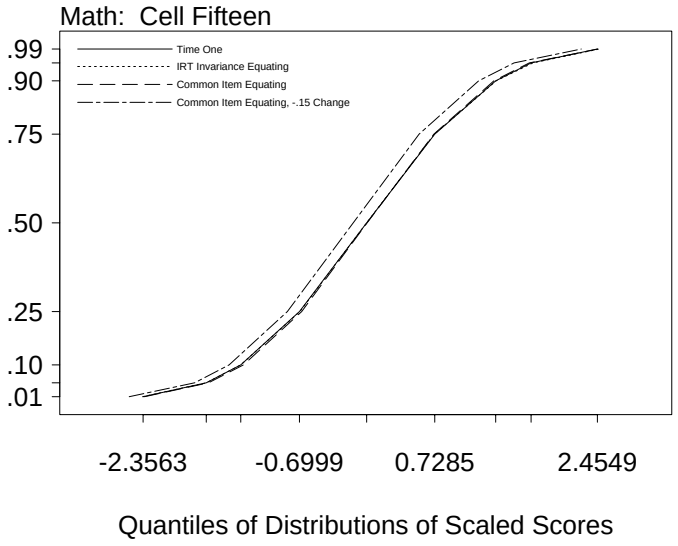


Figure B.38. Quantiles of Scale Distributions for Cell 16 (Mathematics Simulation)

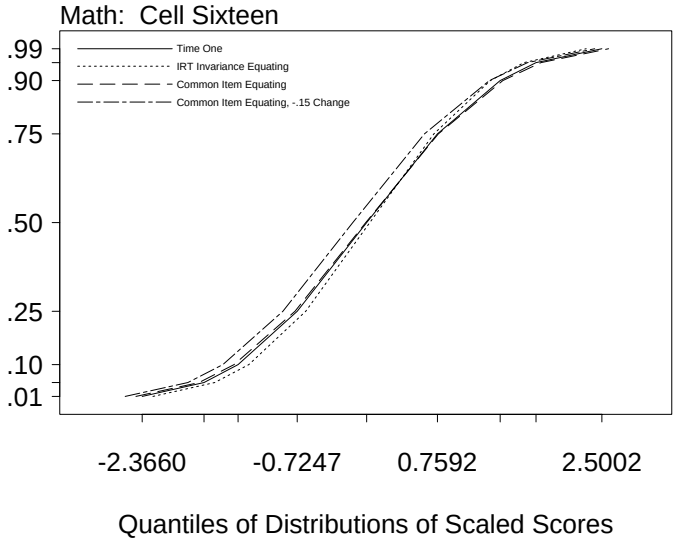


Figure B.39. Quantiles of Scale Distributions for Cell 17 (Mathematics Simulation)

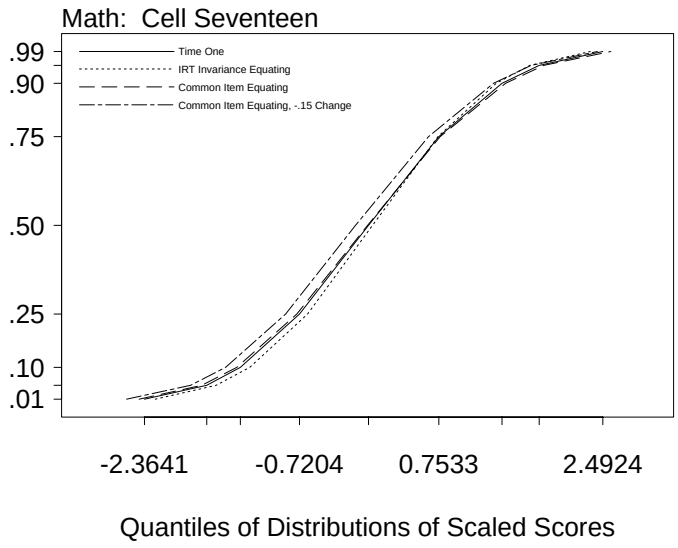


Figure B.40. Quantiles of Scale Distributions for Cell 18 (Mathematics Simulation)

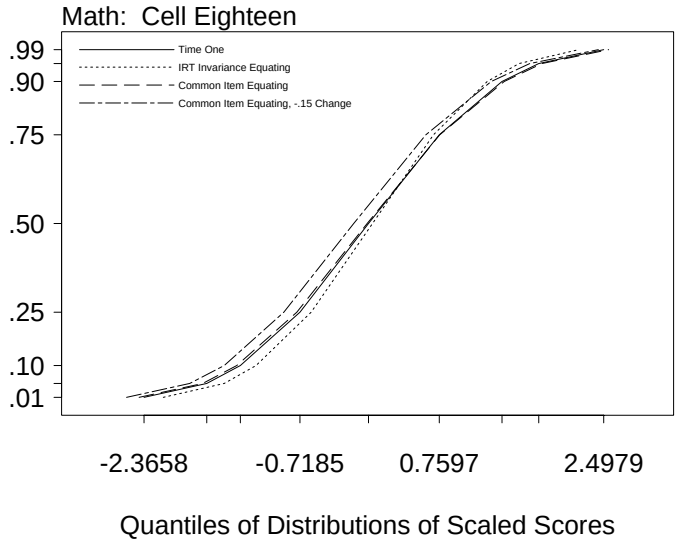


Figure B.41. Quantiles of Scale Distributions for Cell 19 (Mathematics Simulation)

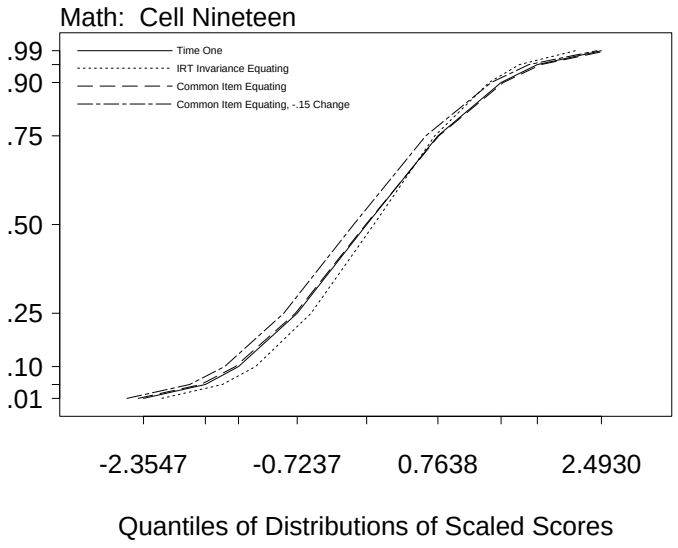
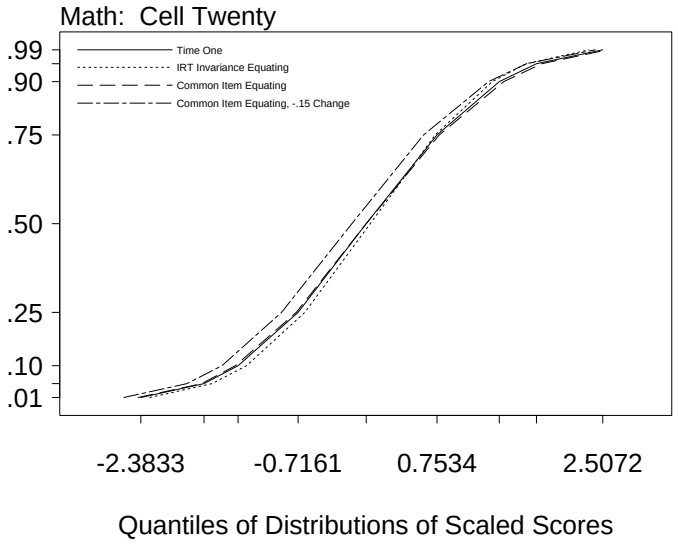


Figure B.42. Quantiles of Scale Distributions for Cell 20 (Mathematics Simulation)



References

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement*, 12, 261-280.

Bock, R.D. & Zimowski, M.F. (1996). Multiple group IRT. In W.J. van der Linden and R. K. Hambleton (Eds.) *Handbook modern of item response theory.* New York: Springer-Verlag.

Johnson, E. G. & Carlson, J.E. (1992). *NAEP technical report.* National Center for Educational Statistics: Washington, DC.

Knuth, D. (1981). *The art of computer programming: Volume 2, seminumerical algorithms (Second Edition).* Addison Wesley: Reading, MA.

Marsaglia, G., & Zaman, A. (1991). A new class of random number generators. *Annals of Applied Probability, 1(3)*, 462-480.

Mazzeo, J. & Donoghue, J. (1995). *Comparing IRT-based equating procedures for trend measurement in a complex test design.* Paper presented at the annual meeting of the American Educational Research Association, New York, 1995.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*, 993-997.

Mislevy, R., and Bock, R.D. (1990). *Bilog 3: Item analysis and test scoring with binary logistic models.* Scientific Software, Inc.: Moorseville, IN.

Stocking, M.L. & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Methods, 7*, 201-210.