Finding Stability: Comparing Methods for Detecting Unstable Item Parameters in IRT Equating

Jeffrey T. Steedle

Paper presented at the 2022 Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Abstract

In IRT-based common item equating, instability in common item parameters can introduce error into IRT scale transformations, subsequent equating results, and, ultimately, examinee scores. This study compared five methods of identifying items with significant parameter drift. Rather than detecting simulated parameter drift like many prior studies, this study used expected equating results as evaluation criteria, which was possible due to the operational use random groups equipercentile equating with an anchor form. Results indicated that two methods produced similarly low equating error while eliminating relatively few items from the common item set. The first was ACT's current practice of flagging items with outlier parameter estimates based on historical distributions. The second was the Delta method, which flags items when transformed proportion correct values are significantly different from expectations.

Introduction

Many testing programs rely upon IRT-based common item nonequivalent groups equating to maintain the meaning of test scores over time (Kolen & Brennan, 2004). The success of this equating design depends partly on a linear scale transformation, and that transformation could be estimated poorly if the common item parameters differ significantly across testing occasions. For that reason, such items are commonly omitted from the estimation of the scale transformation slope and intercept. This general process is sometimes referred to as a "stability check" to detect items with unstable parameters (or parameter drift).

This study compared five methods of detecting unstable item parameters to address the research question, "Which stability check procedure minimizes equating error?" This contrasts with prior studies, which tended to focus on the

accurate identification of simulated item parameter changes. To avoid simulating item parameter changes, which may be unlike real-world changes, this study used expected equating results as evaluation criteria. This was possible because the data came from a test equated via random groups equipercentile equating with an anchor test form administered (and calibrated) at two different times. The identity equating function (0=0, 1=1, 2=2, etc.) served as the criterion when the anchor form was equated to itself via IRT true-score equating. Deviations from the identity function could reflect factors such as random estimation error, sample differences, and common item selection, but holding all else constant, differences in results reflect only the method of removing unstable items. Overall, this research provides practical guidance for operational testing programs using IRT-based common item nonequivalent groups equating.

Background

Several prior studies compared methods of detecting unstable item parameters. For example, Karkee and Choi (2005) observed that four different methods flagged different items and caused discernable differences in students' test scores. Murphy, Little, Fan, Lin, and Kirkpatrick (2010) compared several methods using simulated 3PL data with "realistic" item parameter changes. Results indicated that robust z tended to flag too many items, whereas d^2 (a measure of area between item characteristics curves) and differences in a or b parameters sometimes failed to identify true item parameter changes. Using simulated data, Meyer and Huynh (2010) estimated the Type-I error rate to be .09 to .12 for the robust z method, and they observed that detection power varied with sample size, magnitude of parameter drift, and number of common items. Likewise, Arce and Lau (2011) estimated the robust z Type-I error rate to be .08 to .13. More recently, He and Cui (2020) simulated parameter drift and identified the least absolute values method as the best among five for accurately estimating scale transformation parameters. Rewley and Kaliski (2021) found that d^2 performed relatively well compared to regression residuals across a variety of simulated conditions.

Most prior research focused on the accurate detection of items with simulated unstable parameters. This was sensible, of course, but the equating process does not stop there. The result of greatest practical value is whether the subsequent scale transformation leads to "correct" equating results. This principle guided the design of this study, which used expected equating results as the criteria when comparing five methods of detecting unstable item parameters in operational test data.

Method

Data

The data for this study come from the ACT® test, which is administered nationwide for college admissions, college course placement, and high school accountability (ACT, 2020). The full ACT test consists of four sections: English (75 items), math (60 items), reading (40 items), and science (40 items). Typically, the ACT is equated using random groups equipercentile equating, with one or more previously equated "anchor" forms spiraled with new forms to maintain score scale consistency over time. Though equipercentile equating does not use IRT, a 3PL IRT-calibrated item pool is maintained to support pre-equating for ACT International testing, the PreACT®, and other special testing contexts (e.g., testing with certain accommodations). After calibration, Stocking-Lord scale transformation parameters (Stocking & Lord, 1983) are estimated using two sets of anchor form IRT parameter estimates: (1) when the items were calibrated as a new form, and (2) when the items were calibrated as the anchor form. The transformation is then applied to items parameters from all newly equated forms to put them on the ACT bank IRT scale.

Detecting Outlier Items

Five methods of detecting unstable item parameter estimates were applied in this study. The first— referred to as the "ACT difference" method— flags items for which any difference between 3PL item parameters ($a_{i,y} - a_{i,x}$, $b_{i,y} - b_{i,x}$, or $c_{i,y} - c_{i,x}$) falls outside the corresponding range observed for 95% of items based on historical ACT data. The next method was the Delta (Δ) method (Angoff & Ford, 1973), which was developed as a way to detect differential item functioning (DIF). The method involves transforming proportion correct (p) to normal curve deviates (z) to the Delta scale using $\Delta = 4z + 13$. Items are flagged when their perpendicular distances from the major axis (D_i) fall outside the range of $\pm 1.96 \times s_D$, where s_D is the standard deviation of the perpendicular distances from a line of best fit.

The robust z method entails calculating z statistics for a (discrimination) and b (difficulty) parameter estimate differences using the medians and interquartile ranges (rather than means and standard deviations), which makes z robust to extreme values (Huynh & Meyer, 2010). For example, the following equations are used to calculate the robust z statistic for differences between a parameter estimates for item i.

$$D_{a,i} = \ln(a_{i,y}) - \ln(a_{i,x})$$

$$z_{a,i} = \frac{[D_{a,i} - median(D_a)]}{0.74 \times IQR(D_a)}$$

Items are flagged when z falls outside the range of ± 1.96. The d^2 method is akin to IRT-based DIF analyses, wherein the squared difference between item characteristic curves is calculated and weighted by the distribution of ability (Murphy et al., 2010).

$$d_i^2 = \sum_k \left[P_{ix}(\theta_k) - P_{iy}(\theta_k)\right]^2 g(\theta_k)$$

Items are flagged when d^2 is greater than the 95th percentile of the d^2 distribution based on historical data. The fifth and final method was the $\chi^2_{df=2}$ DIF approach described by Lord (1980), which tests whether the a or b parameter estimates differed significantly. That is,

$$\chi_i^2 = \boldsymbol{v}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{v}_i,$$

where $\boldsymbol{v}_i = [a_{i,x} - a_{i,y}, b_{i,x} - b_{i,y}]$ and $\boldsymbol{\Sigma}_i$ is the corresponding variance-covariance matrix.

Analysis

For this study, the following test forms were equated:

1. Form X with item parameter estimates from the (later) administration during which Form X was the equating anchor

2. Form X with item parameter estimates from the (earlier) administration during which Form X was initially equated

All item parameter estimates were on the scale of the ACT IRT-calibrated item pool. The true equating relationship between raw scores on Form X (anchor) and Form X (initial equating) should be 0=0, 1=1, 2=2,..., J=J. In each of 100 replications, 25% of items were randomly selected to serve as the common item set. Their item parameters and proportions correct were input to the five stability check methods. Then, the resulting Stocking-Lord scale transformation constants (based on the common items with stable parameters) were applied to put all parameters on the same scale, and IRT true-score equating was conducted. Differences between equating results and the identity function indicated bias, and the standard deviation of equating results across replications revealed variation caused by methods of detecting items with unstable parameters.

Results

Results are presented here for one full ACT test (Battery A) that was originally equated in 2018 and used as an anchor form in 2020. Table 1 provides descriptive statistics for the number of items flagged by each detection method. Across test

sections, the ACT difference method tended to flag the fewest items, followed closely by the Delta method, then d^2. Consistent with prior research, robust z tended to flag more common items than other methods, but the Lord DIF method flagged the most items by far (typically about 30–40% of items). This was likely related to the sensitivity of χ^2 statistics to sample size, which was always greater than 2,000 examinees.

Table 1. Descriptive Statistics for Number of Items Flagged for Unstable Item Parameters Across 100 Replications (Battery A)

Section	Statistic	ACT Diff.	Delta	Robust z	d^2	Lord DIF
English (75 items, 19 common)	Mean	0.37	0.61	2.86	1.15	7.28
	Median	0	1	3	1	7
	Minimum	0	0	0	0	2
	Maximum	2	2	8	4	12
Math (60 items, 15 common)	Mean	0.08	0.84	2.04	1.80	4.75
	Median	0	1	2	2	5
	Minimum	0	0	0	0	1
	Maximum	1	2	6	5	9
Reading (40 items, 10 common)	Mean	0.47	0.28	1.80	0.46	4.16
	Median	0	0	1	0	4
	Minimum	0	0	0	0	1
	Maximum	2	1	5	2	7
Science (40 items, 10 common)	Mean	0.00	0.13	1.36	0.24	3.63
	Median	0	0	1	0	3.5
	Minimum	0	0	0	0	0
	Maximum	0	1	5	2	8

To illustrate equating bias, Figure 1 shows the average difference between equated raw scores and the identity function. Differences among the five methods were generally small for the English and science tests, but there was greater variation for the math and reading tests. The differences were summarized with a weighted root mean squared difference (wRMSD) statistic (Harris & Crouse, 1993), with weights equal to proportions of examinees at each raw score in 2020 (the year in which this ACT test form served as the equating anchor). As shown in Table 2, the wRMSD values were generally similar across the five methods with a few exceptions: wRMSD for d^2 was higher for math, wRMSDs for robust z and Lord DIF were higher for reading, and wRMSD for robust z was higher for science.

Figure 2 shows the standard deviation of the equated raw scores for the five detection methods. Across test sections, the variation tended to be greatest for Lord DIF followed by robust z. This result was not surprising considering that these two

methods removed the most and second most items from the common item set (Table 1). Among the other methods, ACT difference often exhibited the least variation, and Delta was nearly as low. The measures of bias and variation were combined to calculate the mean squared error (MSE) at each raw score (MSE = Bias^2 + Variance), and these were weighted to generate the weighted root mean squared error (wRMSE; Table 2). On average across test sections, equating error as measured by wRMSE was lowest for the ACT difference method, followed by Delta, d^2, robust z, and Lord DIF (Table 2).

Figure 1. Average Difference between Equated Raw Scores and the Identity Function Across 100 Replications (Battery A)

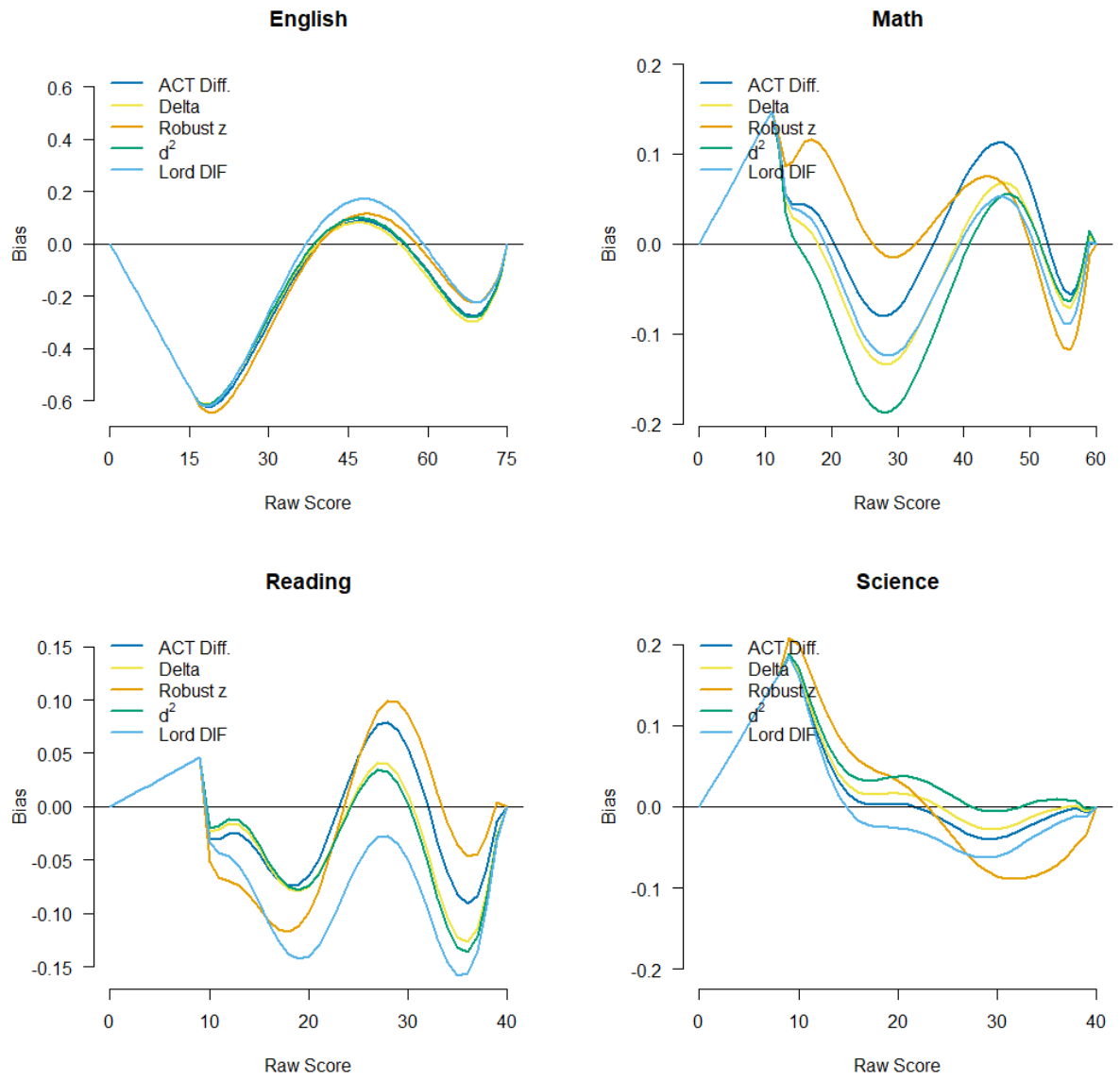
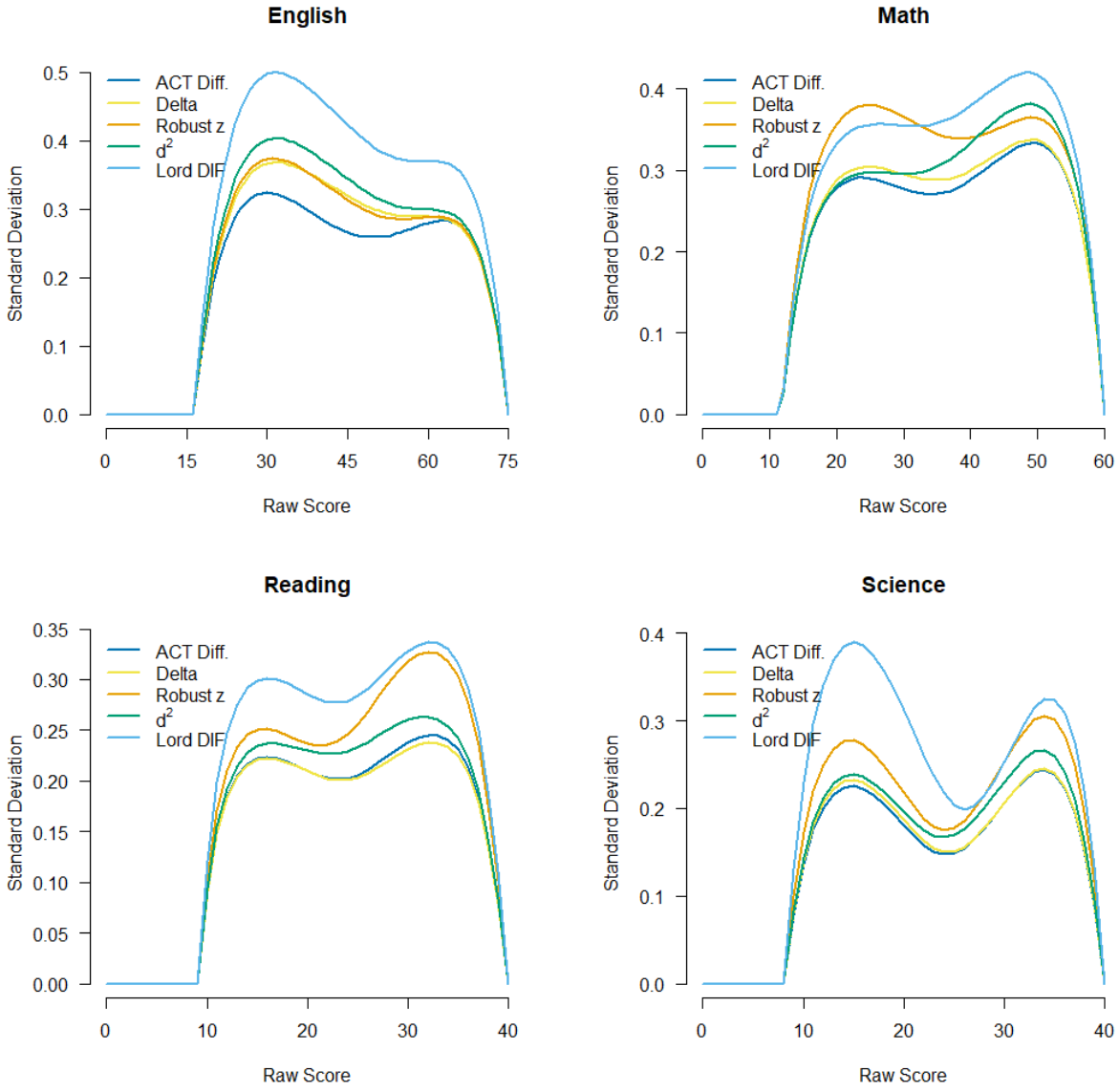


Table 2. Weighted Root Mean Squared Difference and Error (Battery A)

Statistic	Section	ACT Diff.	Delta	Robust *z*	d^2	Lord DIF
Bias/wRMSD	English	0.216	0.215	0.222	0.212	0.211
	Math	0.065	0.078	0.066	0.108	0.073
	Reading	0.056	0.060	0.073	0.062	0.098
	Science	0.049	0.049	0.080	0.053	0.057
wRMSE	English	0.350	0.376	0.380	0.393	0.460
	Math	0.283	0.296	0.345	0.319	0.354
	Reading	0.217	0.215	0.270	0.236	0.303
	Science	0.196	0.200	0.247	0.214	0.295

Figure 2. Standard Deviation of Equated Raw Scores Across 100 Replications (Battery A)



Replications

To assess the generalizability of the results presented above, all analyses were repeated on two additional ACT batteries. Battery B was equated in 2019 and used as the anchor in 2020; Battery C was equated in 2020 and used as the anchor in 2021. Full results for Batteries B and C are provided in the Appendix. Table 3 summarizes results in terms of rankings, with 1 indicating the fewest items flagged, lowest bias/wRMSD, lowest variance, and lowest wRMSE. The general trends in results were quite similar across the three batteries. That is, the ACT difference and Delta methods flagged the fewest items and exhibited the least equating error. Those methods were generally followed by d^2, robust z, and Lord DIF. Note that rankings hide the fact that some methods were nearly tied on some metrics. For example, all five methods exhibited similar levels of bias for Battery B (Appendix Table A2). Indeed, with the exception of Lord DIF, which had particularly high variance due to flagging so many items, each method of detecting significant item parameter drift performed well.

Table 3. Summary of Results

Result	Battery	ACT Diff.	Delta	Robust *z*	d^2	Lord DIF
Items Flagged	A	1	2	4	3	5
	B	2	1	4	3	5
	C	2	1	4	3	5
Bias/wRMSD	A	1	2	5	3	4
	B	2	1	3	5	4
	C	1	4	2	5	3
Variance	A	1	2	4	3	5
	B	1	2	4	3	5
	C	2	1	3	4	5
wRMSE	A	1	2	4	3	5
	B	1	2	4	3	5
	C	1	2	3	4	5

Notes: Results were averaged across test sections to rank the methods for each full ACT test. Ranks for variance were based on a weighted mean variance, with weights equal to proportions of examinees at each raw score.

Conclusions

Using operational data and expected equating results as evaluation criteria, this study compared methods of detecting significant parameter drift among items used in IRT-based common item nonequivalent groups equating. It is assumed that the optimal method exhibits low bias and variance in equating results (i.e., low equating error). Another desirable property— though secondary in importance— is

flagging relatively few items for significant parameter drift because this helps maintain content representation in the common item set. Thus, the ACT difference and Delta methods should be preferred since those methods produced the least biased and least variable equating results, while removing the fewest items from the common item set. The d^2 and robust z methods performed satisfactorily, and their future use should not be discouraged. However, the Lord DIF approach was very sensitive to item parameter differences— likely due to overpowered χ^2 tests— and this caused higher variation in equating results. Overall, results support continued use of the ACT difference method for the ACT test. The Delta method, which is easily implemented, may be preferred for newer testing programs without historical data to set norms for identifying significant item parameter drift.

References

ACT. (2020). *The ACT technical manual*. ACT. Iowa City, IA: ACT. http://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*(2), 95–105. https://doi.org/10.1111/j.1745-3984.1973.tb00787.x

Arce, A. J., & Lau, C. A. (2011, April). *Statistical properties of 3PL robust z: An investigation with real and simulated data sets*. Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, *6*(3), 195–240. https://doi.org/10.1207/s15324818ame0603_3

He, Y., & Cui, Z. (2020). Evaluating robust scale transformation methods with multiple outlying common items under IRT true score equating. *Applied Psychological Measurement*, *44*(4), 296–310. https://doi.org/10.1177/0146621619886050

Huynh, H., & Meyer, P. (2010). Use of robust z in detecting unstable items in item response theory models. *Practical Assessment, Research & Evaluation*, *15*(2), 1–8. https://doi.org/10.7275/ycx6-e864

Karkee, T., & Choi, S. (2005, April). *Impact of eliminating anchor items flagged from statistical criteria on test score classifications in common item equating*.

Annual Meeting of the American Educational Research Association, Montreal, Canada.

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

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

Meyer, P., & Huynh, H. (2010, April). *Evaluation of the robust z procedure for detecting item parameter drift in 3PLM and GPCM mixed format items*. Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Murphy, S., Little, I., Fan, M., Lin, C.-H., & Kirkpatrick, R. (2010, May). *The impact of different anchor stability methods on equating results and student performance*. Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Rewley, K. J., & Kaliski, P. (2021, June). *Comparing the performance of item parameter drift detection methods*. Annual Meeting of the National Council on Measurement in Education, Online.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210. https://doi.org/10.1177/014662168300700208

Appendix

Table A1. Descriptive Statistics for Number of Items Flagged for Unstable Item Parameters Across 100 Replications (Battery B)

Section	Statistic	ACT Diff.	Delta	Robust *z*	d^2	Lord DIF
English (75 items, 19 common)	Mean	1.18	0.59	4.41	1.44	5.45
	Median	1	1	4	1	5
	Minimum	0	0	1	0	1
	Maximum	3	2	10	5	10
Math (60 items, 15 common)	Mean	1.12	0.54	3.51	2.01	5.49
	Median	1	1	3	2	6
	Minimum	0	0	1	0	2
	Maximum	4	2	8	6	9
Reading (40 items, 10 common)	Mean	0.56	0.27	1.95	0.43	3.00
	Median	0.5	0	2	0	3
	Minimum	0	0	0	0	0
	Maximum	2	1	5	2	7
Science (40 items, 10 common)	Mean	0.90	0.08	2.23	0.55	3.99
	Median	1	0	2	0	4
	Minimum	0	0	0	0	1
	Maximum	3	1	5	3	7

Table A2. Weighted Root Mean Squared Difference and Error (Battery B)

Statistic	Section	ACT Diff.	Delta	Robust *z*	d^2	Lord DIF
Bias/wRMSD	English	0.124	0.124	0.144	0.201	0.184
	Math	0.333	0.321	0.323	0.323	0.322
	Reading	0.037	0.037	0.040	0.051	0.059
	Science	0.143	0.143	0.144	0.144	0.151
wRMSE	English	0.331	0.346	0.393	0.367	0.385
	Math	0.465	0.462	0.483	0.478	0.529
	Reading	0.207	0.212	0.231	0.226	0.271
	Science	0.250	0.250	0.272	0.283	0.277

Figure A1. Average Difference between Equated Raw Scores and the Identity Function Across 100 Replications (Battery B)

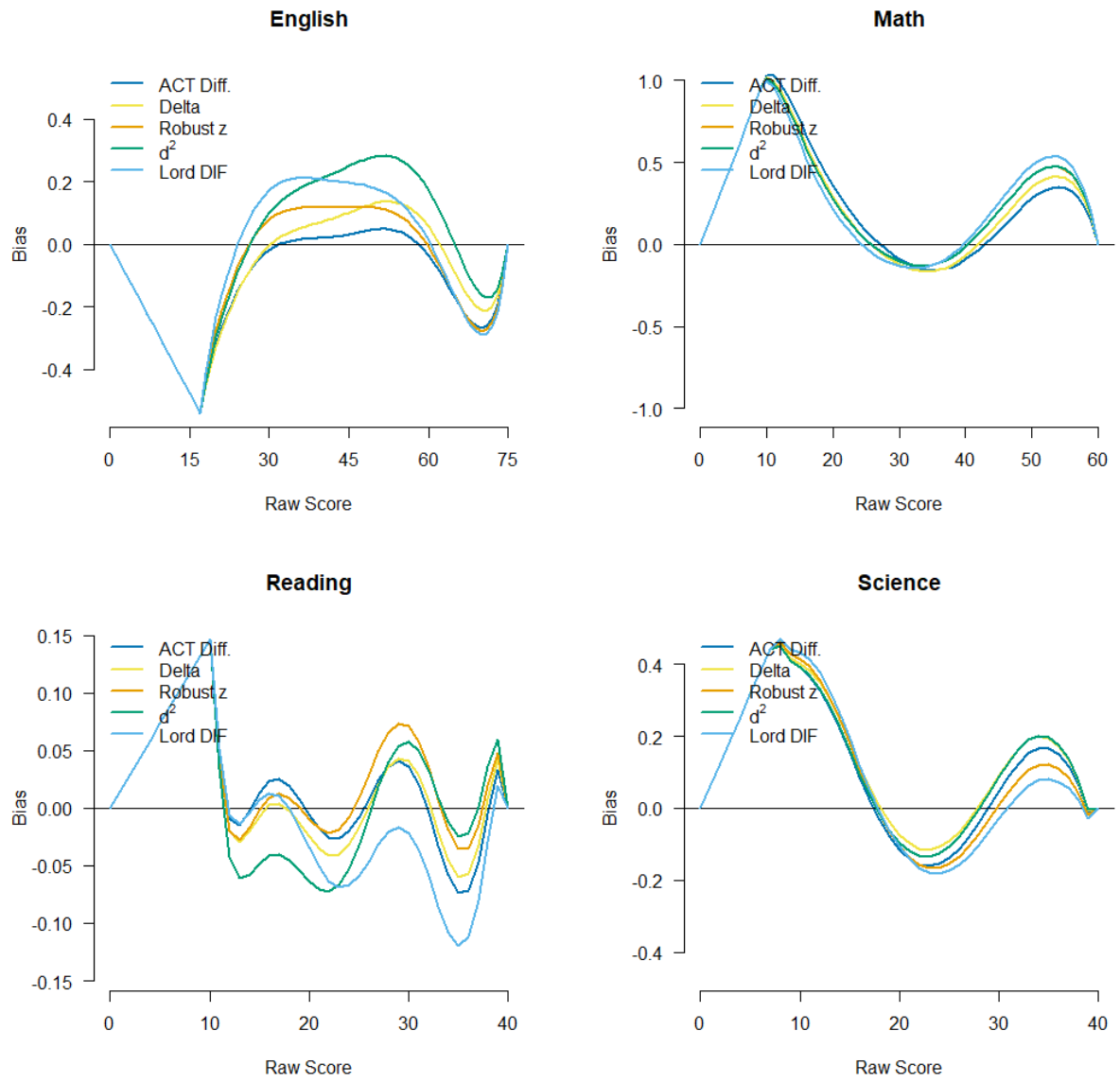


Figure A2. Standard Deviation of Equated Raw Scores Across 100 Replications (Battery B)

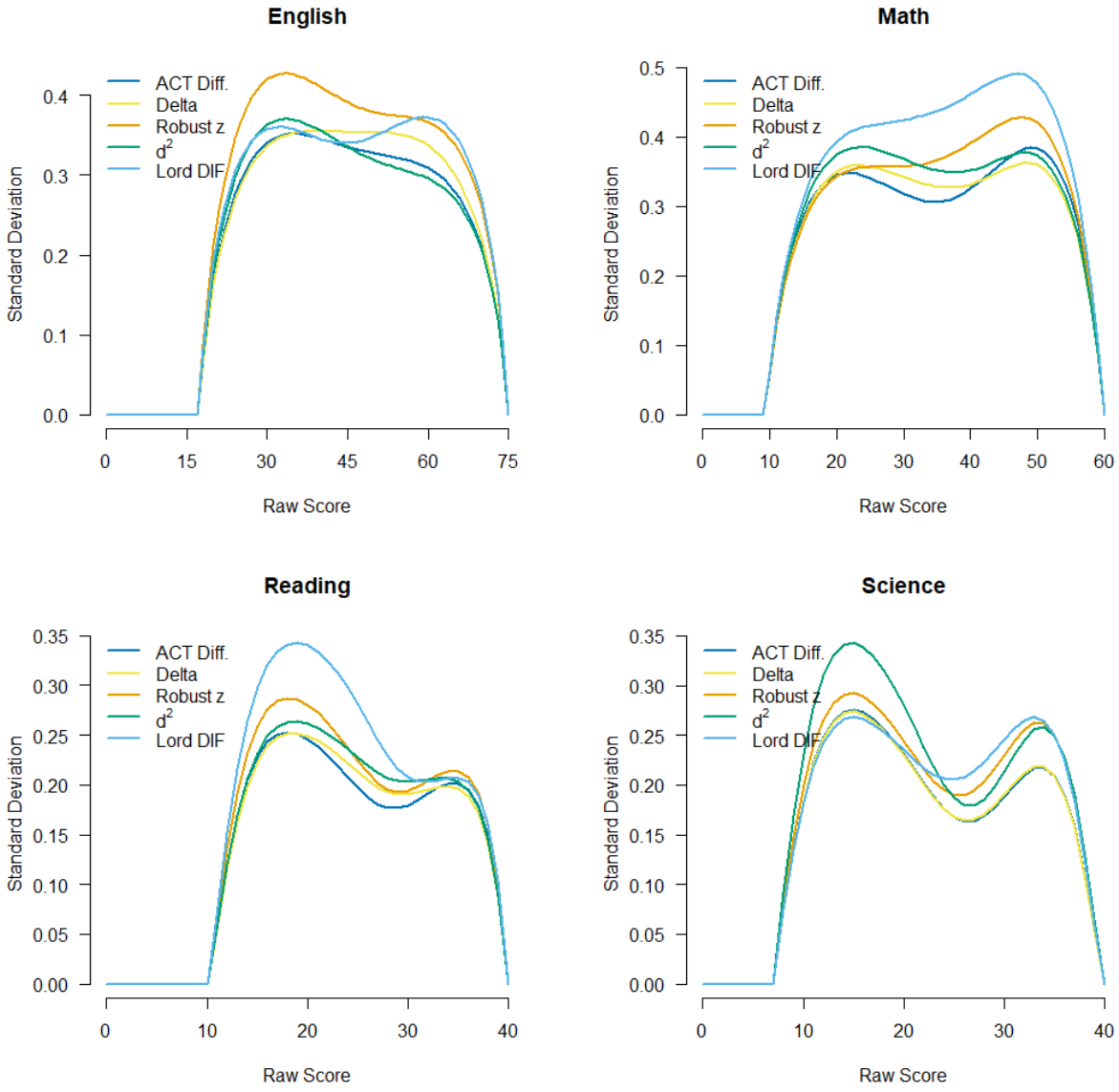


Table A3. Descriptive Statistics for Number of Items Flagged for Unstable Item Parameters Across 100 Replications (Battery C)

Section	Statistic	ACT Diff.	Delta	Robust *z*	d^2	Lord DIF
English (75 items, 19 common)	Mean	1.77	0.58	3.35	1.99	7.02
	Median	2	1	3	2	7
	Minimum	0	0	1	0	2
	Maximum	5	2	8	6	11
Math (60 items, 15 common)	Mean	1.67	0.19	2.50	2.20	5.36
	Median	2	0	2	2	5
	Minimum	0	0	0	0	1
	Maximum	7	1	9	5	9
Reading (40 items, 10 common)	Mean	0.77	0.38	1.86	0.42	2.97
	Median	1	0	2	0	3
	Minimum	0	0	0	0	0
	Maximum	3	1	5	3	6
Science (40 items, 10 common)	Mean	0.80	0.23	1.78	0.60	4.06
	Median	1	0	2	0	4
	Minimum	0	0	0	0	1
	Maximum	3	1	5	2	8

Table A4. Weighted Root Mean Squared Difference and Error (Battery C)

Statistic	Section	ACT Diff.	Delta	Robust *z*	d^2	Lord DIF
Bias/wRMSD	English	0.520	0.519	0.522	0.520	0.523
	Math	0.456	0.446	0.455	0.442	0.448
	Reading	0.124	0.142	0.124	0.133	0.130
	Science	0.361	0.365	0.370	0.379	0.371
wRMSE	English	0.639	0.634	0.677	0.668	0.702
	Math	0.578	0.568	0.574	0.604	0.646
	Reading	0.266	0.281	0.304	0.277	0.340
	Science	0.439	0.440	0.470	0.483	0.542

Figure A3. Average Difference between Equated Raw Scores and the Identity Function Across 100 Replications (Battery C)

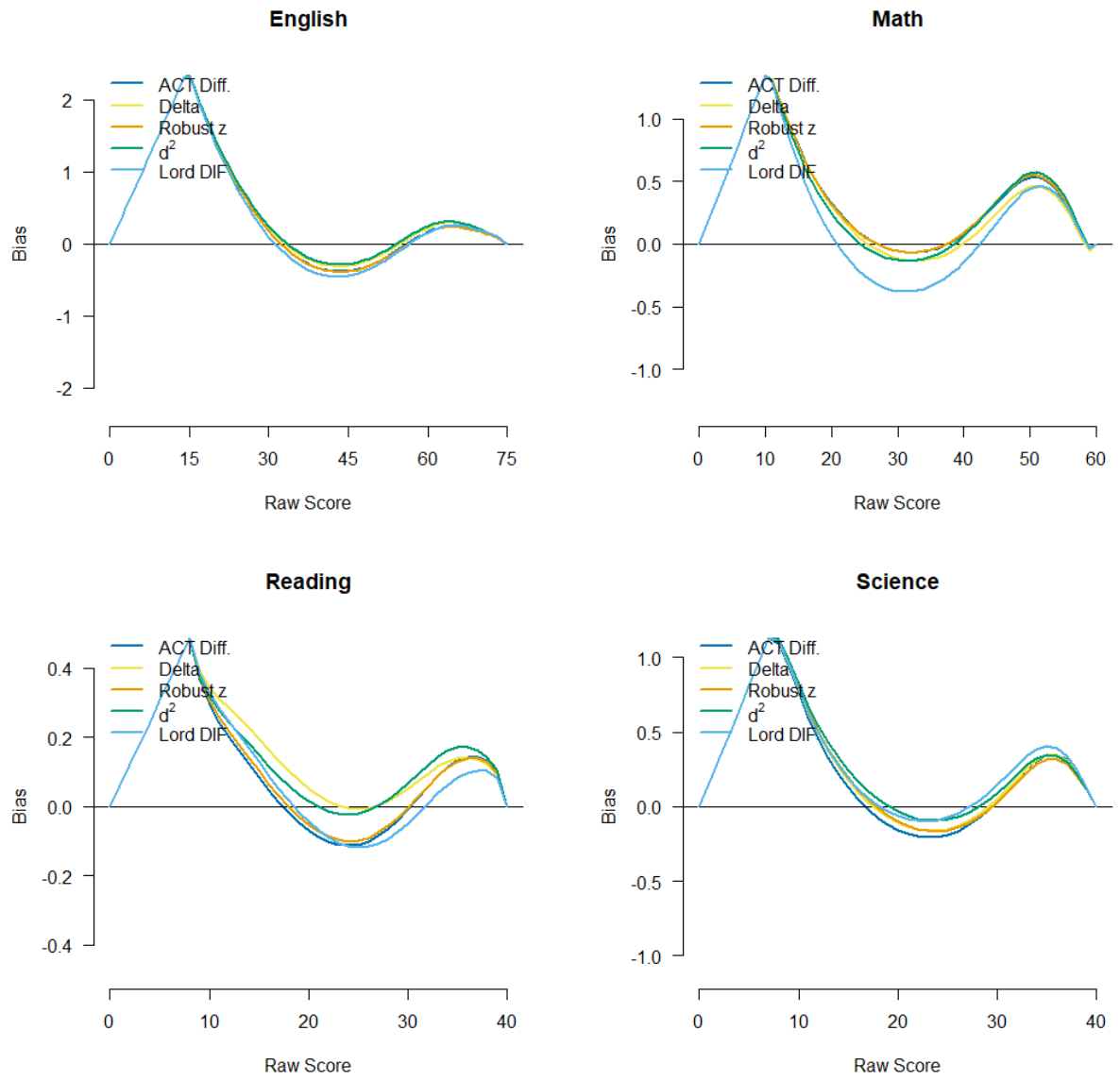


Figure A4. Standard Deviation of Equated Raw Scores Across 100 Replications (Battery C)