A peer reviewed, open-access electronic journal: ISSN 1531-7714

Equating Test Scores with Different Linkage Plans Using IRT Observed-Score Equating

Marie Wiberg, *Umeå University*

Abstract: Linkage plans are used when equating different test versions from different administrations of standardized achievement tests. The overall aim is to examine item response theory (IRT) observed-score equating when different linkage plans have been used when the average ability among test takers in different administrations and/or the average item difficulty in different administrations differ as well as if there are more variation in test scores in some administrations and if a harder anchor test is used. The results indicate that direct equating without an equating chain always yields the smallest bias, root mean squared error and standard errors, and the longest equating chain always yielded the highest errors. More differences could be seen when the test forms and or the item difficulty varied more between groups and test forms. The equated values from using different linkage plans also differed depending on the examined scenario. The results are discussed together with limitations, future research and practical recommendations.

Keywords: NEAT Design, Item-Response Theory Equating, Anchor Test, Item Response Theory, Anchor Test

Introduction

In large-scale standardized testing programs, such as those used for college admissions, multiple test forms are often administered to preserve test security. To ensure comparability of scores across these forms, test equating is required to place them on a common scale (Gonzaléz & Wiberg, 2017). Equating relies on specific data collection designs to establish score relationships between test forms. The equivalent groups (EG) design involves administering different test forms to statistically equivalent populations, while the nonequivalent groups with anchor test (NEAT) design uses a common set of anchor items (i.e., an anchor test) administered alongside each form to adjust for group differences. However, when multiple test forms are used over time or simultaneously, determining how to link them becomes a nontrivial problem. A linkage plan is required to specify the structure of equating relationships across test forms — whether through direct equating to a single reference form, pairwise equating of test forms administered concurrently, or through chain equating across a sequence of test forms. The choice of linkage plan directly affects the stability, fairness, and interpretability of the resulting scores. Thus, beyond the equating design itself, the construction

of an appropriate linkage plan is a critical component in maintaining the validity of longitudinal or large-scale assessments.

Previous research on linkage plans has explored a range of methodological and empirical dimensions. One focus has been the accumulation of equating error resulting from the use of multiple links between test forms or from employing various linking strategies with empirical data (e.g., Guo, 2010; Guo, Liu, Dorans, & Feigenbaum, 2011; Haberman, Guo, Liu, & Dorans, 2008; Liu, Curley, & Low, 2009; Puhan, 2009; Taylor & Lee, 2010). Moses, Deng, and Zhang (2011) investigated the impact of incorporating multiple anchor tests in the linking process. Lee and Ban (2009) evaluated the relative performance of four item response theory (IRT) linking procedures within an EG design, considering three sampling designs, two sample size levels, and two item count levels. Livingston and Antal (2010), using a NEAT design and empirical data, examined alternative linkage plans and proposed adjustments to the final equating when multiple linkage paths are employed. Additionally, analytical studies have addressed both random and systematic errors in multiple equatings (Haberman & Dorans, 2011) and in equating chains (Haberman, 2010).

Battauz (2013) proposed the use of complex linkage plans within the NEAT design, using IRT equating with both empirical data and simulation studies. Her work introduced a method for averaging equating coefficients across multiple linking paths and provided asymptotic standard errors for both indirect and averaged equating coefficients. Wiberg (2017) conducted a comparative analysis of traditional equating methods, kernel equating, and IRT observed-score kernel equating, applying two distinct linkage plans to real test data. More recently, Wiberg (2021) investigated the performance of two linkage plans using frequency estimation, chained equating, and kernel equating methods within both EG and NEAT designs, utilizing data from a college admissions test alongside simulated datasets. The findings consistently indicated that different linkage plans yield varying equated scores and standard errors across both EG and NEAT designs, with notable differences observed among the equating methods. However, a key limitation of these studies was the restriction to only examine two linkage plans. A general conclusion emerging from all these research studies is that linkage plan selection significantly influences the equated scores, although the specific factors contributing to these variations remain insufficiently understood.

The present study diverges from prior research in several important aspects. First, earlier studies have primary focused on traditional equating methods or, as in Wiberg (2021), kernel equating techniques. Note that Liu, Guo, and Dorans (2014) explored various linkage plans and concluded that equating based on a single old test form resulted in the greatest scale drift, which tended to increase over time. More stable conversions with reduced variation were achieved when equating to two or three old test forms. However, their study did not utilize IRT observed-score equating and was primarily concerned with the number of old test forms used rather than the structure of the linkage plan itself. Second, most previous investigations have relied exclusively on either empirical or simulated data to evaluate different linkage plans. In contrast, the current study employs both empirical and simulated data to assess the performance of IRT observed-score equating across multiple linkage configurations. Third, we are using IRT observed-score equating with a NEAT design, which is a combination that has not been studied before. Although Lee and Ban (2009) examined IRT-based linkage procedures, their focus was limited to the EG design, excluding the NEAT design. Also, Haberman (2009) used IRT parameter estimates from a large number of test forms linked pairwise. While conceptually relevant due to its use of IRT, his work did not empirically investigate different linkage plans using either real or simulated data. Fourth, as we are using IRT observed-score equating we investigated several conditions that may influence the equated values, including variations in ability distributions across time points, differences in item difficulty and item discrimination, the difficulty of the anchor test, and if there were larger variation in score distributions among test groups. Although some of these conditions—such as more able test groups and more difficult test forms—have been explored in the context of linkage plans using traditional and kernel equating methods (Wiberg, 2021), they have not been examined in the context of linkage plan with IRT observed-score equating. Fifth, whereas most studies have

considered only two different linkage plans (e.g., Wiberg, 2021) although Liu, Gao and Dorans (2014) is an exception but for EG design and a traditional equating method, our study adopts a broader scope by evaluating four distinct linkage plans in the NEAT design with IRT observed-score equating.

In summary, to the best of our knowledge, no previous study has systematically examined the use of IRT observed-score equating with a NEAT design and varying the test takers' ability and the item characteristics in the test forms across multiple linkage plans.

The overall aim of this study is to investigate the performance of IRT observed-score equating under various linkage plans, particularly in contexts where differences exist across test administrations in terms of average test taker ability, item difficulty, item discrimination, score variability, and anchor test difficulty. To evaluate equating performance, we examine equated scores, bias, root mean squared error (RMSE), and standard errors. Furthermore, the linkage plans in the empirical study is also discussed in light of the four linkage plan rules outlined by Kolen and Brennan (2014) on p. 298, which recommend: (1) minimizing the number of links that influence score comparisons across successive administrations (i.e., avoiding equating strain); (2) linking test forms administered at the same time of year whenever possible; (3) minimizing the number of links connecting each test form to the initial form; and (4) avoid linking back to the same test form too often.

The remainder of the paper is organized as follows. The next section introduces the IRT observed-score equating methodology, followed by a small-scale empirical study involving four different linkage plans. Subsequently, a simulation study is presented, incorporating scenarios in which average ability levels and item characteristics vary across administrations. The paper concludes with a discussion of the findings, including limitations and directions for future research.

IRT Observed-Score Equating

IRT observed-score equating (Lord, 1980) uses marginal score distributions and an IRT model to define conditional score probabilities. These marginal distributions are obtained by assuming a latent ability distribution and summing the conditional probabilities across all ability levels. Once the marginal score distributions for both test forms are established, equipercentile equating is applied to equate the score scales from the two test forms. Assume a new test form X with test scores X and an old test form Y with test scores Y. Assume further that X and Y are continuous and let $f(x|\theta)$, $f(y|\theta)$ and $\psi(\theta)$ be the conditional scores and ability θ distributions. Denote their cumulative density functions (CDFs) with $F_x(x)$ and $F_y(y)$, respectively. These CDFs are obtained from cumulating $f(x) = \int f(x|\theta)\psi(\theta)d\theta$ and $f(y) = \int f(y|\theta)\psi(\theta)d\theta$. The IRT-observed score equating transformation from X to Y is then defined as

$$\varphi_Y(x) = F_Y^{-1}(F_X(X)) \quad (1)$$

The conditional scores distributions are typically obtained using a recursive algorithm proposed in Lord and Wingersky (1984), although other alternatives exist (González, Wiberg, & von Davier, 2016).

When a NEAT design is employed and item parameters have been calibrated separately for each test form, it is necessary to transform the IRT scales to a common metric prior to conducting IRT observed-score equating. Four commonly used scale transformation methods include mean-mean, mean-sigma, Haebara, and Stocking-Lord. Kim and Kolen (2007) found that the Haebara (1980) and Stocking-Lord (1983) methods yield comparable results, while Lee and Ban (2010) reported that the Haebara method may offer slightly improved performance. Accordingly, this study adopts the Haebara method in conjunction with the two-parameter logistic (2PL) IRT model throughout. The Haebara (1980) method seeks to minimize

$$H_{crit} = \Sigma_i \Sigma_j [\pi_{ij}(\theta_i, a_j, b_j) - \pi_{ij}\left(\theta_i, \frac{a_j}{A}, Ab_j + B\right)]$$

where $\pi_{ij}(\theta_i, a_j, b_j)$ is the probability to answer an item *j* correctly by test taker *i* with item discrimination a_j and item difficulty b_j, and A and B are the equating coefficients.

Empirical Study

Real test data were used to illustrate four linkage plans across four administrations of the Swedish Scholastic Aptitude Test (SweSAT), a college admissions examination administered biannually—once in the spring (designated as A) and once in the fall (designated as B). The SweSAT is a multiple-choice, paper-and-pencil test comprising two sections: quantitative and verbal, each containing 80 dichotomously scored items. These sections are equated separately. In the present study, only the quantitative section was analyzed. This section assesses test takers' abilities in areas such as data sufficiency, mathematics, quantitative comparisons, and interpretation of diagrams, tables, and maps. It is divided into two subsections of 40 items each, both constructed according to the same test specifications.

An external 40-item anchor test, constructed according to the same test specifications as the two quantitative subtests of the SweSAT, was administered to a smaller subsample of test takers, while the remaining participants received 40 pretest items. A NEAT design was employed, consistent with the operational equating procedures used for the SweSAT. Since the anchor test form is distributed across different test centers during each administration, it is generally unlikely that repeat test takers encounter the same anchor form more than once. In this study, no individual received the anchor test form on multiple occasions. During the time of these administrations, the SweSAT test score was valid for applying to higher education during five years, and the test takers can repeat the SweSAT as many times as they want, as only the highest score is used when applying to college. Test takers who take the same season administration is assumed to be more similar than those who take different season administrations (i.e. A and B).

The four test administrations were equated using four distinct linkage plans, all employing the Haebara method. In the first linkage plan (labelled 41), test form 4B was directly equated to test form 1B. The second plan (labelled 421) involved a two-step linkage, where test form 4B was first equated to test form 2B, which was subsequently equated to test form 1B. Similarly, the third linkage plan (labelled 431) equated test form 4B to test form 3A, followed by an equating to test form 1B.

In the final linkage plan (labelled 4321), test form 4B was sequentially equated to test form 3A, then to test form 2B, and finally to test form 1B. This approach violates the three first linkage plan rules. Notably, the first rule—minimizing the number of links that influence score comparisons across successive administrations—is not explicitly examined in this study, as doing so would require the evaluation of more complex linkage plans. A typical violation of the first rule would be to equate several fall and spring administrations separately and only use a single bridge link between one spring and one fall administration, thereby increasing the number of indirect connections and potential sources of error. Linkage plan 41, which employs a single direct link between test forms 4B and 1B, adheres to all established linkage rules. In contrast, linkage plans 431 and 4321 violate the second rule by incorporating administrations from different seasons — specifically, both spring (A) and fall (B) — which introduces potential variability due to seasonal differences among test takers. Additionally, linkage plans 421, 431, and 4321 breach the third rule by utilizing more steps than necessary to establish the connection to the base form, thereby increasing the complexity and potential error in the equating process. None of the four linkage plans violate the fourth rule, because our focus is to identify the most suitable equating plan when linking back to the oldest test form, not to implement multiple linkage plans that link back to different old test forms.

The empirical study was done in R and IRT observed-score equating was performed with the R package `equateIRT` (Battauz, 2015). The test score distributions were plotted to be able to compare them. The equated values and the standard errors (SE) when using the four linkage plans were compared when using IRT observed-score equating. The SEs were obtained from `equateIRT` which are using the SE formula from Ogasawaara (2003).

Results of the Empirical Study

Descriptive statistics for the four quantitative test forms administered across four SweSAT administrations, as well as for the common external anchor test form administered during each session, are presented in Table 1. As shown, the number of test takers varied substantially across administrations, ranging from 40,431 to 74,437. This fluctuation is likely influenced by broader socioeconomic factors, particularly changes in unemployment rates and labor market conditions in Sweden.

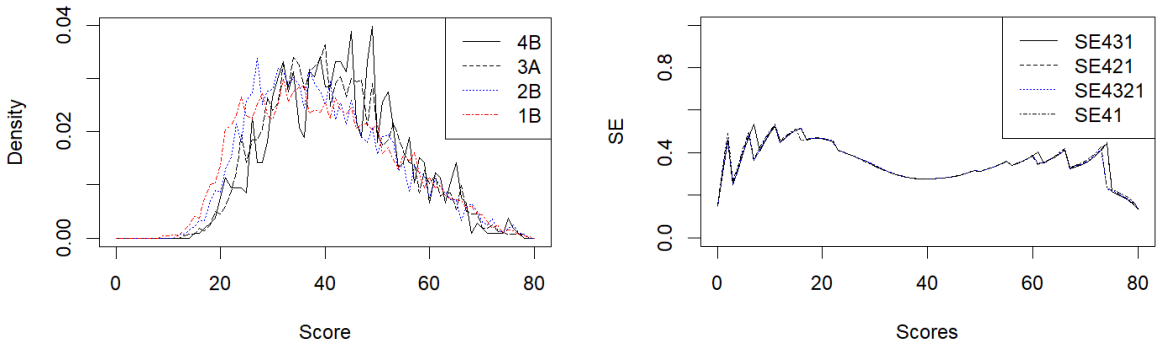
Table 1. Mean and Standard Deviations of the Total Scores and Anchor Scores Together with Number of Test Takers for the SweSAT Administrations

	Total scores			Anchor scores		
Adm	*M*	*SD*	*N*	*M*	*SD*	*N*
1B	37.91	13.43	40,431	18.40	6.55	5,263
2B	42.52	13.31	58,840	16.64	6.62	2,783
3A	43.00	12.35	74,437	16.71	6.44	2,826
4B	42.90	12.54	60,008	17.37	6.11	1,052

Note. Adm = Administration

Historically, higher unemployment tends to correlate with increased SweSAT participation, as individuals seek alternative pathways to higher education. Table 1 also reveals that the mean scores on the anchor test varied across administrations. This variation may similarly reflect shifts in the test-taking population's average ability, with larger cohorts—often driven by economic pressures—typically exhibiting lower average performance. Despite these differences in mean scores, the standard deviation of anchor test scores remained relatively stable across all four administrations, suggesting consistent score dispersion regardless of cohort size.

Figure 1. Score distributions for the four administrations and standard errors (SE) for the four linkage plans



As illustrated in the left panel of Figure 1, the score distributions across the four administrations were highly similar. Consequently, the equated scores (not shown) and the standard errors (SE)—depicted in the right panel of Figure 1—were also comparable across the four linkage plans. This outcome aligns with

expectations: when score distributions are closely matched, different linkage strategies tend to yield similar equating results. However, given that score distributions may vary across administrations in practice, we conducted a simulation study to examine how differences in average test taker ability, average item difficulty, and average item discrimination impact equated values when using different linkage plans.

Simulation Study

The 2PL IRT model was used to generate item responses with item parameters set to $a \sim U(0.5,2)$ and $b \sim N(0,1)$ for the baseline case, which is like Andersson & Wiberg (2017). In the baseline case (Case 0) the four test group populations', V (administration 1), P (administration 2), Q (administration 3) and Z (administration 4), ability θ was assumed to follow the standard normal distribution. Population size was set to 40,000 in line with the smallest administration size in the empirical study. Likewise, sample size of test takers for the anchor test administration was set to N = 2,000, which is similar in size to two of the anchor test administrations in the real data empirical study. For the simulation study, we used n = 40 items for the regular test and 20 items for the anchor test, as that is a common test size for standardized tests and the SweSAT quantitative test is given as two sets of 40 items that are added together. Also, an anchor test form should range from 20 to 60 items (e.g., Kolen & Brennan, 2014, p. 271; Petersen et al., 1982). By choosing 20 anchor items and 40 regular items we had the same proportion of anchor items with respect to the regular items as in the empirical study. We used IRT observed-score equating and the Haebara method with four different linkage plans in 39 different cases, which can be seen in Table 2, where the cases are grouped as six (0-5) overall scenarios.

In the baseline case (Case 0), all administrations were similar in difficulty, and the test groups were of equal average ability. The baseline case is thus similar in that sense to the empirical study. In Case 1, all test taker groups were assumed to be of the same average ability, but the average item difficulty was higher in some of the administrations. In Case 2, the average item difficulty is the same in all administrations, but some of the test taker groups have higher average ability (θ+0.5). In Case 3, the test taker groups have on average higher ability (θ+0.5), and the average item difficulty is higher (b+0.5) in some of the administrations. In Case 4, some of Cases 0-3 are repeated but with larger average standard deviations in the ability (θ+0.5) for some of the test taker groups. In Case 5, the anchor test contains on average more difficult items (b+0.5) for some of the 0-4 Cases. In a few of the cases (1f, 2f, 3f, and 4f) we also examined the impact of on average higher item discrimination (a+0.5) in two of the administrations, as that can impact the equated values (e.g., Wiberg, 2016). All scenarios in Table 2 were examined with four different linkage plans and replicated 500 times. Although we examined 39 scenarios in six overall cases, we will only display some of the scenarios in this section which we think are interesting to discuss. The omitted figures can be found on the following github: https://github.com/MarieWiberg/Linkage-plan. The displayed cases are marked in grey in Table 2.

Table 2. Changes in Mean and Standard Deviation in Ability for Populations (V,P,Q,Z), and Changes in Mean Item Difficulty (b) and Item Discrimination (a) in the Test Forms (1,2,3,4)

	M (ability)				*B*				*a (1,2,3,4)*	*SD*
Case	V	P	Q	Z	1(V)	2(P)	3(Q)	4(Z)		
0	0	0	0	0	0	0	0	0		
1a	0	0	0	0	0	0.5	0	0.5		
1b	0	0	0	0	0	0	0	0.5		
1c	0	0	0	0	0	0.5	0.5	0		
1d	0	0	0	0	0	0	0.5	0.5		
1e	0	0	0	0	0.5	0.5	0.5	0		
1f	0	0	0	0	0	0	0	0	0,0,0.5,0.5	
2a	0	0.5	0	0.5	0	0	0	0		
2b	0	0	0	0.5	0	0	0	0		
2c	0	0.5	0.5	0	0	0	0	0		
2d	0	0	0.5	0.5	0	0	0	0		
2e	0.5	0.5	0.5	0	0	0	0	0		
2f	0	0	0.5	0.5	0	0	0	0	0,0,0.5,0.5	
3a	0	0.5	0	0.5	0	0.5	0	0.5		
3b	0	0	0	0.5	0	0.5	0	0.5		
3c	0	0.5	0.5	0	0	0.5	0	0.5		
3d	0	0	0.5	0.5	0	0.5	0	0.5		
3e	0.5	0.5	0.5	0	0	0.5	0	0.5		
3f	0	0	0.5	0.5	0	0.5	0	0.5	0,0,0.5,0.5	
4a	0	0	0	0	0	0	0	0		1.5,1,1,1.5
4b	0	0	0	0	0	0.5	0	0.5		1,1.5,1,1.5
4c	0	0.5	0	0.5	0	0	0	0		1,1.5,1,1.5
4d	0	0	0	0.5	0	0	0	0		1,1.5,1,1.5
4e	0	0.5	0.5	0	0	0	0	0		1,1.5,1,1.5
4f	0	0.5	0	0.5	0	0.5	0	0.5	0,0,0.5,0.5	1,1.5,1,1.5
4g	0	0.5	0	0.5	0.5	0	0.5	0		1,1.5,1,1.5
4h	0	0.5	0	0.5	0	0.5	0	0.5		1,1.5,1,1.5
More difficult anchor +0.5										
50A	0	0	0	0	0	0	0	0		0
51aA	0	0	0	0	0	0.5	0	0.5		0
52aA	0	0.5	0	0.5	0	0	0	0		0
53aA	0	0.5	0	0.5	0	0.5	0	0.5		0
54aA	0	0	0	0	0	0	0	0		1.5,1,1,1.5
54bA	0	0	0	0	0	0.5	0	0.5		1,1.5,1,1.5
54cA	0	0.5	0	0.5	0	0	0	0		1,1.5,1,1.5
54dA	0.5	0.5	0.5	0	0	0.5	0	0.5		1,1.5,1,1.5
54d2A	0	0	0	0.5	0	0	0	0		1,1.5,1,1.5
54eA	0	0.5	0.5	0	0	0	0	0		1,1.5,1,1.5
54fA	0	0.5	0	0.5	0	0.5	0	0.5		1,1.5,1,1.5
54gA	0	0.5	0	0.5	0.5	0	0.5	0		1,1.5,1,1.5

Note. Populations V, P, Q and Z take test forms 1, 2, 3, and 4 respectively.

Evaluation Measures

To evaluate the equating performance using the different linkage plans, we used bias, RMSE, and SE. The estimated equated scores were compared with the estimated true equated scores at each test score. Let xi denote a specific test score for score values $i = 0,\ldots,n$. As noted by Wiberg and González (2016), the definition of SE and standard error of equating (SEE) coincide and SE can be obtained from $SE\big(\hat{\varphi}_Y(x_i)\big) = \sqrt{\text{Var}(\hat{\varphi}_Y(x_i))}$, where $\hat{\varphi}_Y(x_i)$is the estimated equated score for test score x_i. The estimated equated score for replication r=1,..,R is denoted $\hat{\varphi}_Y^{(r)}(x_i)$. We used the true item parameters to obtain the estimated true equating score, which we denote as $\varphi_Y(x_i)$ in the equations below. Using these definitions, the bias was defined as

$$\text{bias}\big(\varphi_Y(x_i)\big) = \frac{1}{R}\Sigma_{r=1}^{R}\left(\varphi_Y^{(r)}(x_i) - \varphi_Y(x_i)\right),$$

and RMSE was defined as

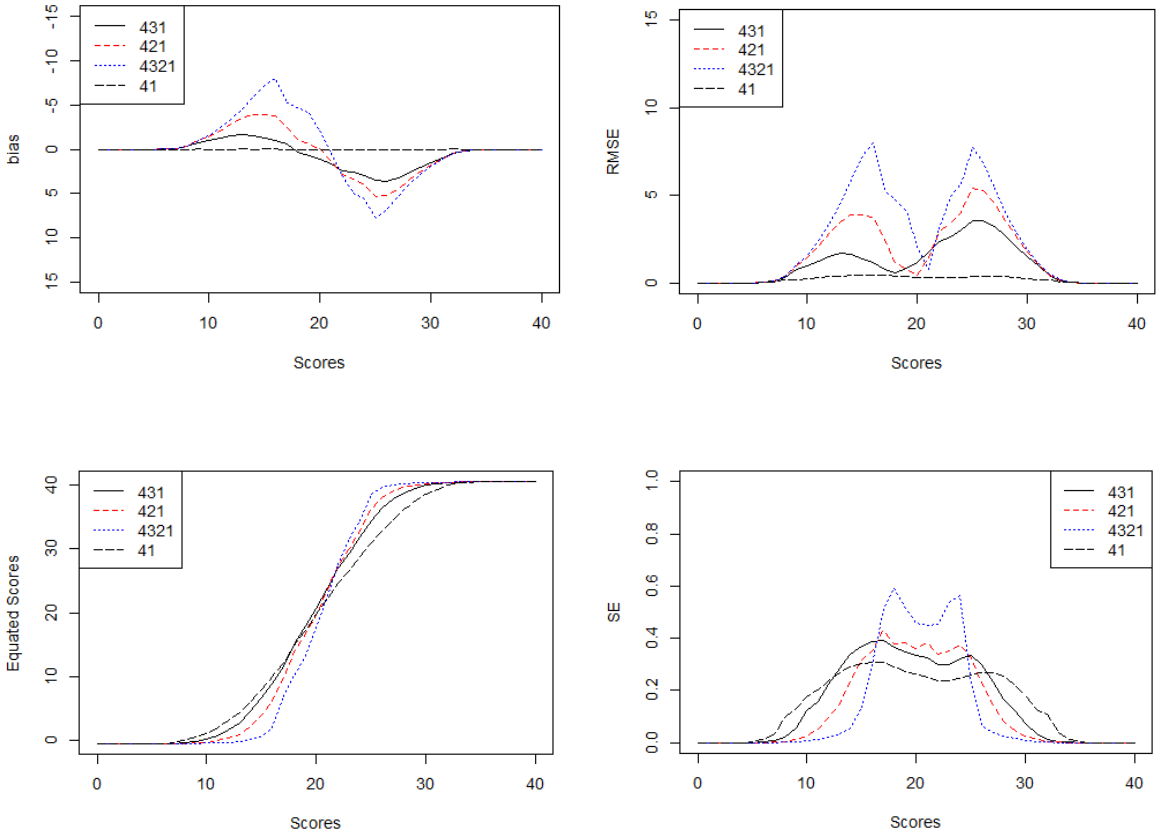
$$\text{RMSE}\big(\varphi_Y(x_i)\big) = \sqrt{\frac{1}{R}\Sigma_{r=1}^{R}(\hat{\varphi}_Y^{(r)}(x_i) - \varphi_Y(x_i))^2}$$

For a discussion of these measures in observed-score equating refer to Chapter 7 in Wiberg and González (2016). The simulation study was done in R and IRT observed-score equating was performed with the R package equateIRT (Battauz, 2015).

Results from the Simulation Study

In general, differences in equated values were observed among the four linkage plans in most of the examined cases. Across all cases, bias was consistently largest for linkage plan 4321 and smallest for linkage plan 41. Since the bias figures appeared quite similar, only the bias and RMSE figures for the baseline case (top of Figure 2) are presented in the paper.

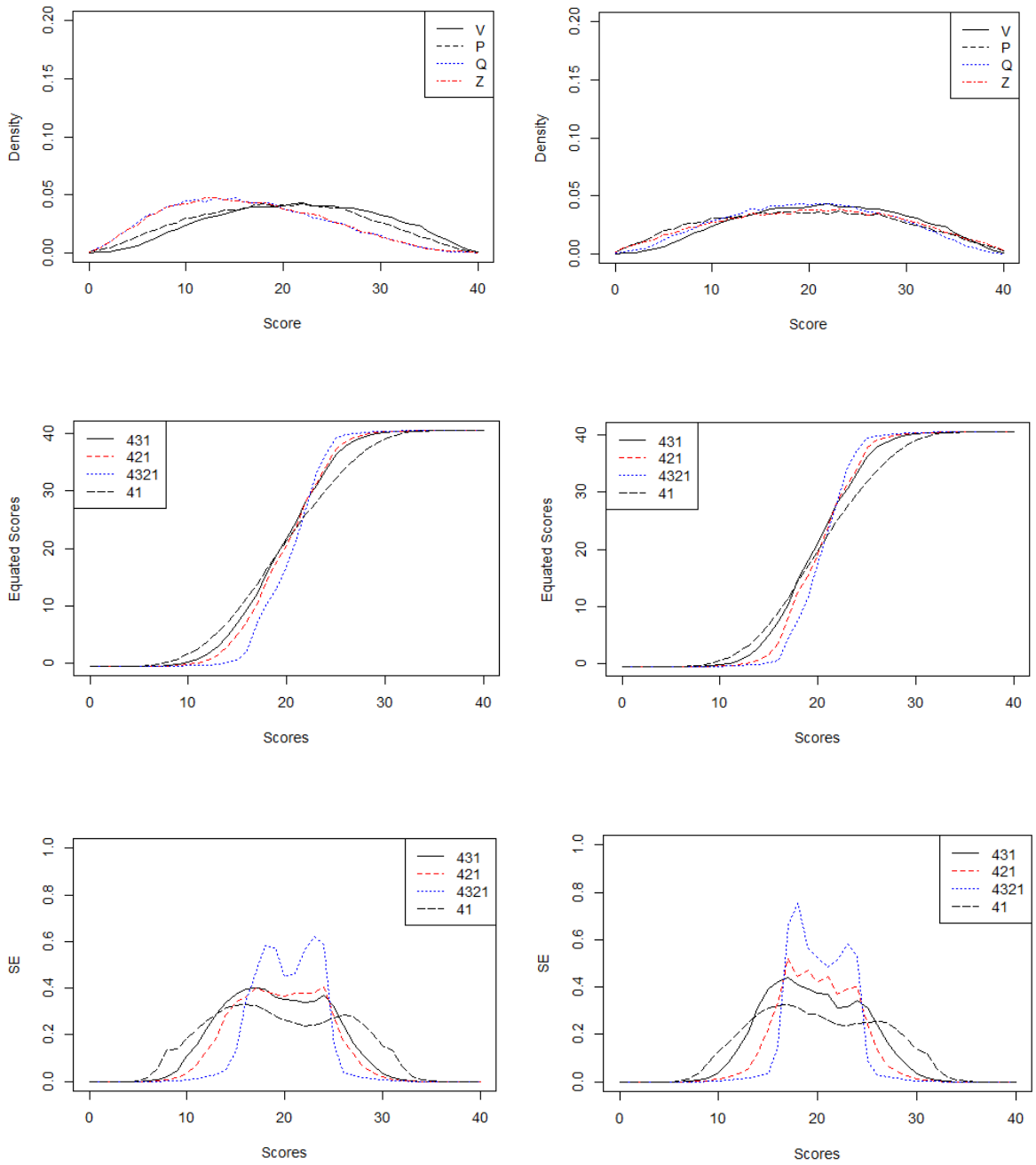
Figure 2. For the baseline case (Case 0), bias and RMSE (top) and equating transformation and SE (bottom)



Since the score distributions in the baseline case were identical across all administrations—due to all ability groups having the same ability level and all administrations sharing the same average item difficulty and item discrimination —we omitted that figure. According to Figure 2, linkage plan 41 exhibited the lowest overall bias and RMSE compared to the other three linkage plans. SEs were generally lowest for linkage plan 41, although it varied across the score scale. As expected, linkage plan 4321 performed the worst in terms of bias, RMSE, and SE for mid-range score values, followed by linkage plans 421 and 431.

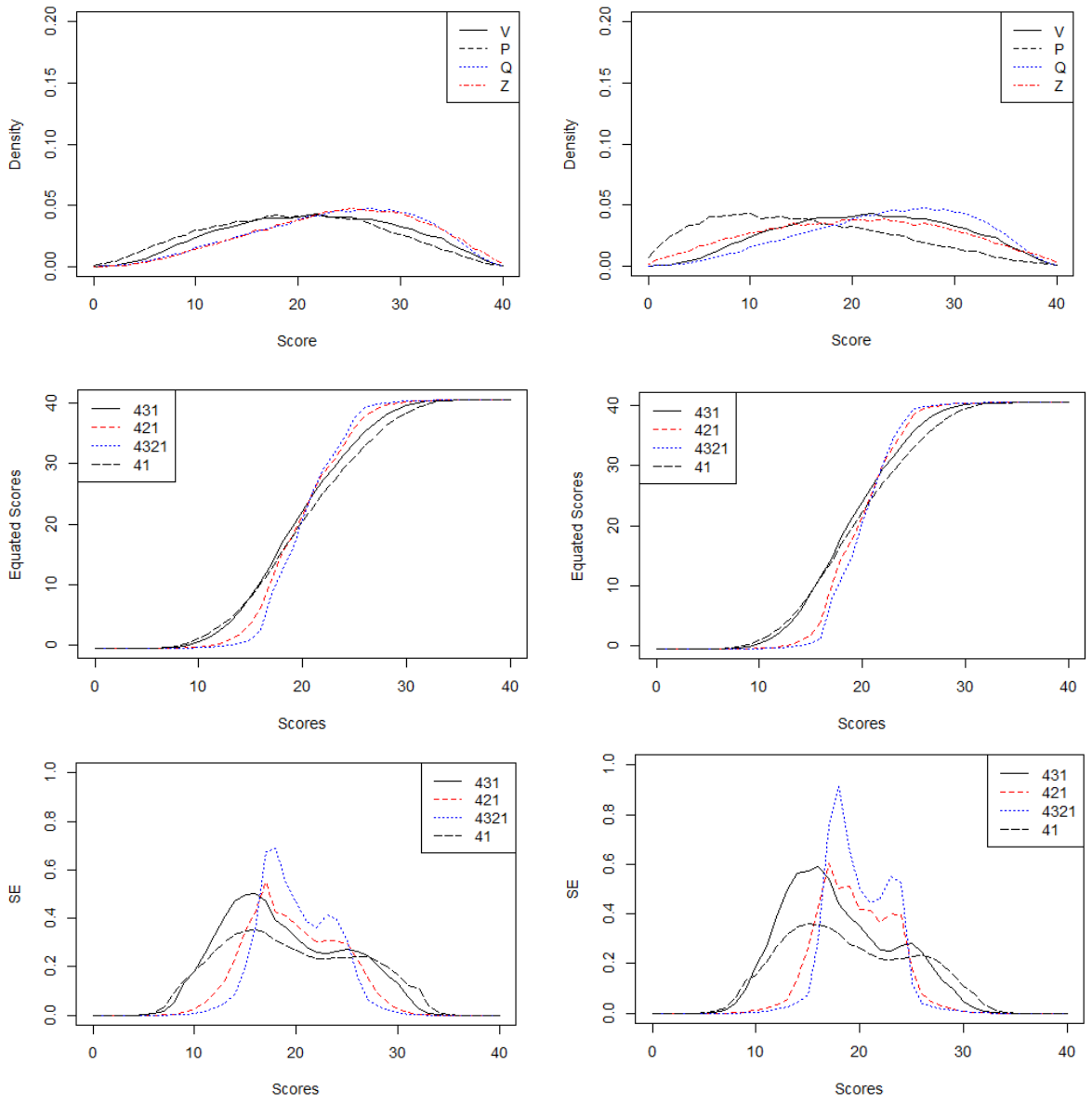
When the average ability was the same across groups, but the average item difficulty differed (i.e., all Cases 1a-1e), the linkage plans produced noticeably different equated values, particularly linkage plan 41 compared to the others. As shown in the left part of Figure 3 for Case 1d, the SEs were notably different for the longest linkage plan, 4321. When item discrimination varied (right part of Figure 3), the equated values differed in a similar manner to when item difficulty was varied, but the SEs were slightly larger for several score values, especially for the longer linkage plan. Although the figures are not shown here, as expected when the two mid test score distributions (test forms 2 and 3) were similar, the SE and the equated scores were also similar for linkage plans 431 and 421 (e.g., Case 1c and Case 1e).

Figure 3. More difficult test forms are given to populations Q and Z (Case 1d) to the left, and more discriminating items are given to populations P and Z (Case 1f) to the right.



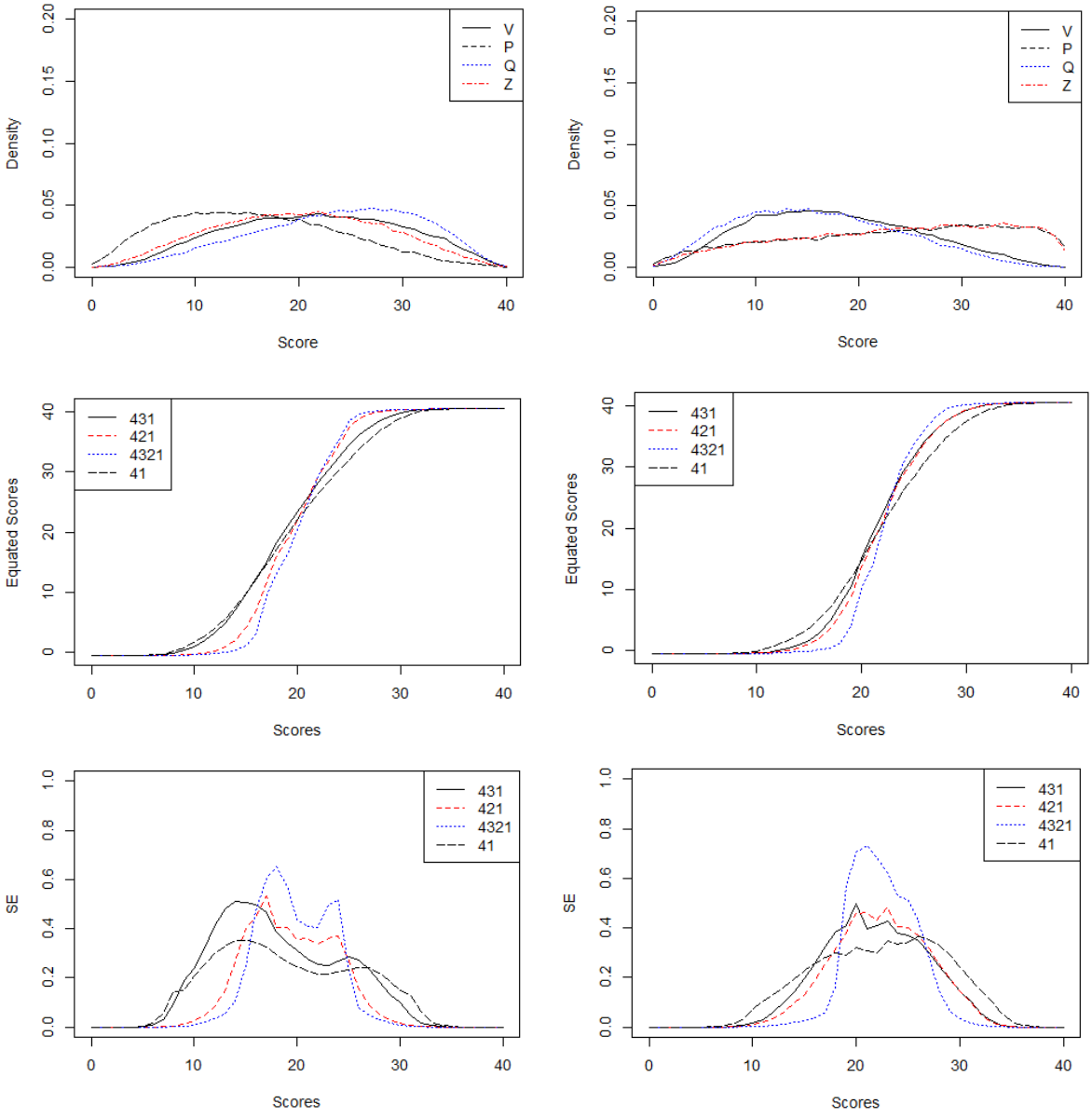
The left part of Figure 4 (Case 2d) illustrates when the average item difficulty remains the same, but the average ability is higher in some groups (i.e., all Cases 2a-2e). The score distributions clearly differ, as do the equated values across the linkage plans. For most score values, the SEs were again lowest for linkage plan 41, followed by 431, 421, and with 4321 performing the worst. The bias and RMSE patterns for Cases 1 and 2 were similar to those in the baseline case and were therefore omitted.

Figure 4. Higher average ability in populations P and Z (Case 2d) to the left, and higher average ability in populations Q and Z together with more discriminating and more difficult items given to populations Q and Z (Case 3f) to the right.



The right part of Figure 4 depicts a scenario in which populations P and Z received more difficult and more discriminating items, while populations Q and Z had higher average ability. In this case, both the SEs and equated values varied across linkage plans. SE was lowest for linkage plan 41 for most scores, although the distributions were more skewed for linkage plans 431 and 41. Linkage plan 4321 had the highest SE in the mid-score range, and since the bias and RMSE resembled those in the baseline case, they were omitted.

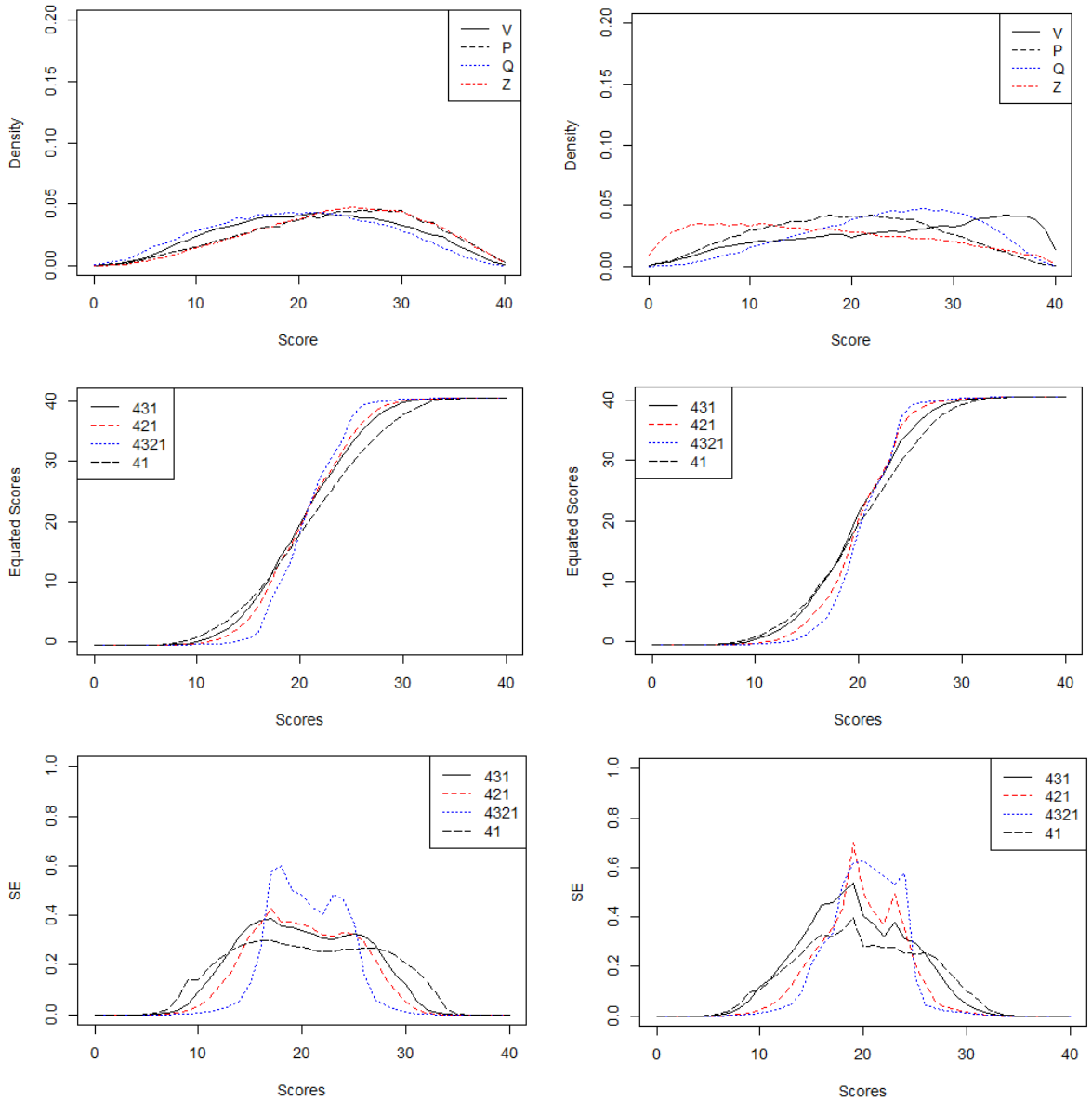
Figure 5. On average more difficult test forms in PZ and more able test groups in QZ (Case 3d) to the left, and more difficult test forms VQ and more able test groups PZ when SD more varied (Case 4g) to the right.



In the various scenarios of Case 3, we examined situations where some administrations had, on average, more difficult items, while others had more able test takers. The left part of Figure 5 (Case 3d) illustrates when more difficult test forms were administered to populations P and Z, and more able test groups were found in populations Q and Z. In general, for Case 3, only small differences in the magnitude of bias and RMSE were observed, with bias and RMSE being slightly higher for linkage plan 4321 and slightly lower for linkage plan 431 compared to the baseline case. While the overall conclusions regarding SE were consistent with previous findings, the shape of the SE curves were similar to Case 2d (more able test takers). The right side of Figure 5 shows Case 4g, where more difficult items were given to populations V and Q, and populations P and Z had on average more able test takers but with greater variation in ability SD. The test score distributions clearly differed between administrations and not surprisingly, these differences led to variations in the equating transformations depending on the linkage plan used.

In Case 5 (Figure 6), we examined scenarios where the anchor test was, on average, more difficult. The left part of Figure 6 illustrates Case 52aA, in which the average ability was higher in some administrations, while the average item difficulty remained the same across all administrations. The right part of Figure 6 shows Case 54dA, where both the average ability and average item difficulty were higher in certain administrations, and the SD of test takers' ability was more varied in two of the administrations. In both cases, the equated scores varied depending on the linkage plan used. As in previous cases, linkage plan 41 consistently yielded the lowest SE across most of the score range, followed by linkage plans 431 and 421, with linkage plan 4321 consistently producing the highest SEs. For the other scenarios in Case 5, the SE curves varied slightly depending on which administration had, on average, more difficult test forms or more able test takers. However, the overall conclusion remained consistent. The bias and RMSE patterns were similar in appearance to those in the baseline case, although the magnitude of bias was slightly smaller.

Figure 6. A more difficult anchor test. In the left panel, average ability was higher in PZ (Case 52aA). To the right, average ability was higher in VPQ, average item difficulty was higher in PZ, and average standard deviation of the test takers' ability was higher in PZ (Case 54dA).



Discussion

The overall aim was to examine different linkage plans when IRT observed-score equating was used, and the average ability among test takers in different administrations and/or the average item difficulty or item discrimination differed, as well as when there was a larger SD in the test takers' ability, or when a harder anchor test was used. Not surprisingly, using different linkage plans only resulted in small changes in the equated values when test forms and test groups were similar. In line with this, if the test score distributions were similar for the middle administrations (test forms 2 and 3), one could use either linkage plan 431 or 421, regardless of variation in item difficulty or item discrimination in the test forms, or variation in ability among the groups. However, this was not the case when the linkage plans differed in length (e.g., comparing linkage plan 4321 with either linkage plans 431 or 421).

Using the longest linkage plan (4321) consistently yielded the highest SEs and RMSEs, regardless of the scenario examined. In general, using a longer linkage plan was always worse in terms of bias, RMSE, and SEs in the mid-score range compared to using a shorter linkage plan. This pattern was also true when there was more variation in test takers' ability or when a more difficult anchor test was used. The equated values from the four linkage plans differed more when the test forms varied more in item difficulty and the test groups differed in ability—a result consistent with Wiberg (2021), who examined different linkage plans in traditional and kernel equating methods. Interestingly, and not previously studied, when item discrimination varied, the equated values across the different linkage plans differed in a similar way to when item difficulty varied, although SEs were slightly larger for several score values, especially for the longer linkage plans. This has practical implications, as it suggests that we should aim to keep not only item difficulty but also item discrimination similar between test forms if we want consistent equated values and SEs.

There were several limitations in this study that should be addressed in future research. Although we examined several scenarios, we could not explore all scenarios of interest. In future research, it would be valuable to vary the test length (e.g., Lee & Ban, 2009; Albano & Wiberg, 2019), to use more than one anchor test (e.g., Moses, Deng, and Zhang, 2011), or to vary the anchor test length, as all these factors are known to affect the equating transformation and could potentially influence the performance of different linkage plans. It is especially important to acknowledge the limitation of using the same anchor test form across all administrations, as was done in this study. Specifically, repeated use of a single anchor test form increases the risk that test takers may become familiar with its content in advance, potentially compromising the validity and reliability of the resulting test scores. One limitation of the overall conclusion—that using a shorter linkage plan is better than using a longer one — is that, in practice, relying on only one test form may threaten test security. The overall conclusion also differs from the empirical study by Liu, Guo, and Dorans (2014), who found that equatings were more stable and showed less variation when more than one old test form was used. A major limitation of their study, however, was that they used only one empirical dataset, which meant they could not control specific conditions. Additionally, their study differs from ours in that we used IRT observed-score equating, while they used a traditional equating method. As we reached different conclusions, future research should examine different linkage plans when equating transformations are averaged (e.g., Holland & Strawderman, 2011), or when equating transformations need to be merged, as discussed by Livingston and Antal (2010).

We used IRT observed-score equating with the Haebara method, but future research could also examine whether using the Stocking-Lord method affects the results, although previous studies have suggested that these methods tend to yield similar outcomes (e.g., Kim & Kolen, 2007). Furthermore, while this study focused on IRT observed-score equating, future work could explore IRT observed-score kernel equating (Andersson & Wiberg, 2017), which allows for item modeling using an IRT model in the pre-smoothing step while benefiting from the flexibility of kernel equating (Wiberg, González & von Davier, 2025). Although initial work has been done with empirical data (Wiberg, 2017) a more structured analysis of the

use of different linkage plans in different scenarios has not yet been performed. This would be particularly interesting, as Wiberg (2021) studied linkage plans using kernel equating but not IRT observed-score kernel equating. Additionally, future research could investigate linkage plans with simultaneous linking.

In practice, we recommend using the least complex linkage plan possible, as this consistently resulted in the smallest bias, RMSE, and SE in the mid-score range across all the scenarios examined. However, we want to emphasize that more complex linkage plans, and the use of multiple anchor test forms may be necessary to avoid compromising test security. We recommend using IRT observed-score equating, as it produced similar results when test forms and groups were comparable and yielded more stable results compared with traditional methods, as examined in Wiberg (2021). It also allows us to incorporate item difficulty and item discrimination into the model, both of which were shown to have an impact on the equated values in the different linkage plans.

Funding
This research was funded by the Swedish Wallenberg MMW 2019.0129 grant and the Swedish Research Council grant 2022–02046.

Received: 4/8/2025. **Accepted:** 11/17/2025. **Published:** 12/5/2025.

Citation: Wiberg, M. (2025). Equating test scores with different linkage plans using IRT observed-score equating. *Practical Assessment, Research, & Evaluation*, 30(1)(11). Available online: https://doi.org/10.7275/pare.3056

Corresponding Author: Marie Wiberg, Department of Statistics, USBE, Umeå University, SE-901 87 Umeå, Sweden. Email: marie.wiberg@umu.se

References

Albano, A. & Wiberg, M. (2019). Linking with external covariates: examining accuracy by anchor type, test length, and sample size. *Applied Psychological Measurement. 43*(8), 597–610. https://doi.org/10.1177/0146621618824855

Andersson, B. & Wiberg, M. (2017). Item response theory observed-score kernel equating. *Psychometrika. 82*(1), 48–66. https://doi.org/10.1007/s11336-016-9528-7

Battauz M. (2013). IRT test equating in complex linkage plans. *Psychometrika.* Jul;78(3): 464–80. https://doi.org/10.1007/s11336-012-9316-y

Battauz M (2015). `equateIRT`: An R Package for IRT Test Equating. *Journal of Statistical Software*, *68*(7), 1–22. https://doi.org/10.18637/jss.v068.i07

González, J., & Wiberg, M. (2017). *Applying test equating methods – using R.* Chapter 1, Cham, Switzerland: Springer. DOI: 10.1007/978-3-319-51824-4

González, J., Wiberg, M., & von Davier A. A. (2016). A note on the Poisson's binomial distribution in item response theory. *Applied Psychological Measurement, 40*(4), 302–310. https://doi.org/10.1177/0146621616629380

Guo, H. (2010). Accumulative equating error after a chain of linear equatings. *Psychometrika*, *75*(3), 438–453. https://doi.org/10.1007/s11336-010-9160-x

Guo, H., Liu, J., Dorans, N. J., & Feigenbaum, M. (2011). *Multiple linking in equating and random scale drift* (ETS Research Report 11–46). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02282.x

Haberman, S. J. (2009). *Linking parameter estimates derived from an item response model through separate calibrations* (ETS Research Rep. 09-39). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02197.x

Haberman, S. J. (2010). *Limits on the accuracy of linking* (ETS Research Report No. 10–22). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2010.tb02229.x

Haberman, S. J., & Dorans, N. J. (2011). *Sources of score scale inconsistency* (ETS Research Report 11–10). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02246.x

Haberman, S. J., Guo, H., Liu, J., & Dorans, N. J. (2008). *Consistency of SAT I: Reasoning test score conversions* (ETS Research Report 08–67). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02153.x

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 144–149. https://doi.org/10.4992/psycholres1954.22.144

Holland, P. W. & Strawderman, W. E. (2011). How to average equating functions, if you must. In A. A. von Davier *Statistical models for test equating, scaling, and linking.* Chapter 6, pp 109–122. Springer.

Kim, S., & Kolen, M. J. (2007). Effects on scale linking of different definitions of criterion functions for the IRT characteristic curve methods. *Journal of Educational and Behavioral Statistics*, *32*(4), 371–397. https://doi.org/10.3102/1076998607302632

Kolen, M. & R. Brennan (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer-Verlag.

Lee, W. C., & Ban, J. C. (2009). A comparison of IRT linking procedures. *Applied Measurement in Education*, *23*(1), 23-48. https://doi.org/10.1080/08957340903423537

Liu, J., Curley, E., & Low, A. (2009). *A scale drift study* (ETS Research Report 09–43). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02200.x

Liu, J., Guo, H. & Dorans, N. (2014). *A comparison of raw-to-scale conversion consistency between single- and multiple-linking using nonequivalent groups anchor test design.* (ETS Research Report No. 14-13). Educational Testing Service. https://doi.org/10.1002/ets2.12014

Livingston, S. A., & Antal, J. (2010). A case of inconsistent equatings: How the man with four watches decides what time it is. *Applied Measurement in Education*, *23*(1), 49–62. https://doi.org/10.1080/08957340903423578

Moses, T., Deng, W., & Zhang, Y. (2011). Two approaches for using multiple anchors in NEAT equating: A description and demonstration. *Applied Psychological Measurement*, *35*(5), 362–379. https://doi.org/10.1177/014662161140551

Ogasawara, H. (2003). Asymptotic standard errors of IRT observed-score equating methods. *Psychometrika*, *68*(2), 193–211. https://doi.org/10.1007/BF02294797

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland & D. B. Rubin (Eds.), Test equating (pp. 71–135). Academic Press.

Puhan, G. (2009). Detecting and correcting scale drift in test equating: An illustration from a large scale testing program. *Applied Measurement in Education*, *22*(1), 79–103. https://doi.org/10.1080/08957340802558391

Stocking, M., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 10*(3), 201–210. https://doi.org/10.1177/014662168300700208

Taylor, C. S., & Lee, Y. (2010). Stability of Rasch scales over time. *Applied Measurement in Education*, *23*(1), 87–113. https://psycnet.apa.org/doi/10.1080/08957340903423701

Wiberg, M. (2016). Alternative linear IRT observed-score equating methods. *Applied Psychological Measurement, 40*(3), 180–199. https://doi.org/10.1177/0146621615605089

Wiberg, M. & González, J. (2016). Statistical assessment of estimated transformations in observed-score equating, *Journal of Educational Measurement*, *53*(1), 106-125. https://doi.org/10.1111/jedm.12103

Wiberg, M. (2017). *Ensuring test quality over time by monitoring the equating transformation*. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W-C. Wang. (Eds.) Quantitative Psychology, Springer. 239–252.

Wiberg, M. (2021). On the use of different linkage plans with different observed-score equipercentile equating methods. *Practical Assessment, Research and Evaluation, 26*(24) 1–16. https://scholarworks.umass.edu/pare/vol26/iss1/23/

Wiberg, M., González, J. & von Davier, A. A. (2025). *Generalized kernel equating with applications in R.* CRC Press.

Wiberg, M., van der Linden, W.J., & von Davier, A. (2014). Local kernel observed-score equating. *Journal of Educational Measurement, 51*(1), 57–74. https://doi.org/10.1111/jedm.12034