Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 57

Comparison of Simultaneous Linking and Separate Calibration with Stocking-Lord Method

Guangyun Liu[†]
Hyung Jin Kim
Won-Chan Lee
YoungKoung Kim

November 2024

[†] Guangyun Liu is Graduate Research Assistant, Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa (email: lguangyun@uiowa.edu). Hyung Jin Kim is Research Scientist, Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa (email: hyungjinkim@uiowa.edu). Won-Chan Lee is Director, Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa (email: won-chan-lee@uiowa.edu). YoungKoung Kim is Senior Psychometrician at the College Board (email: ykim@collegeboard.org).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: https://education.uiowa.edu/centers/casma

Contents

List of Tables

List of Figures

Acknowledgements

This research was funded by a contract with the College Board. We extend our gratitude to the College Board for their financial support. Additionally, we would like to thank all the individuals and teams at the College Board who contributed their expertise and insights throughout this project. Their invaluable input and collaboration were crucial to the completion of this report.

Abstract

This study evaluates the relative performance of simultaneous linking and Stocking-Lord's scale transformation under two linking scenarios. Simultaneous linking is hypothesized to reduce linking error by circumventing long linking chains. Our findings indicate that the performance of these two linking methods is highly context-dependent. Despite trivial differences in overall performance, simultaneous linking shows superior recovery of item discrimination parameters, whereas Stocking-Lord's method is more effective in recovering item difficulty parameters.

1 Introduction

Testing programs that utilize Item Response Theory (IRT) often encounter the challenge of locating item parameters from different test forms on a common scale. To address this issue, IRT linking procedures can be employed to ensure consistent resolution of linear indeterminacy across the various forms. Of available methods, linking via separate IRT calibrations for each form is often preferred, as it does not require data from every form to be available at the same time. This procedure entails the separate estimation of item parameters for each form and then putting them on the common scale through a linear transformation.

Two most popular approaches for scale transformation are the Haebara method (Haebara, 1980) and the Stocking-Lord method (Stocking & Lord, 1983). These two methods typically involve two forms, new and old forms, and use characteristic curves to find linear transformation coefficients, a slope (A) and an intercept (B), which are then used to place item parameters for one form on the scale of the other form. Scale transformation is commonly conducted from new form to old form. The Haebara method finds the transformation coefficients that minimize the sum of the squared differences between item characteristic curves for the common items in the two forms. On the other hand, the Stocking-Lord method uses test characteristic curves in the minimization process. Kolen and Brennan (2014) provide further discussion on the differences between these two methods.

When multiple test forms are linked on a common scale, the process often involves multiple linkages that accumulates errors over time. Concurrent calibration can link multiple test forms through a single calibration by combining the new and old forms. In concurrent calibration, items not taken by other groups are treated as missing or not reached (Lord, 1980). However, concurrent calibration requires that all datasets be available at the time of calibration, and the calibration can be a time-consuming process.

Concurrent calibration can potentially minimize "linking error", as it requires only one model specification to estimate all parameters (Briggs & Weeks, 2009). However, concurrent calibration may introduce bias when data are multidimensional. In contrast, separate calibration is less affected by violations of unidimensionality (Béguin & Hanson, 2001). Similar to separate calibration, where multiple sets of common item parameter estimates are obtained over time, concurrent calibration also results in a new set of parameter estimates if one of the forms used has already been calibrated. This scenario presents a challenging decision-making process regarding whether to replace or retain the existing parameter estimates, regardless of which method is used (Lee & Lee, 2018).

1.1 Haberman's Simultaneous Linking

Haberman (2009) introduced a regression-based linking method that is considered a generalization of the log-mean/mean method (Mislevy & Bock, 1990). This method is commonly referred to as simultaneous linking, and as the name implies, it has the capability to link multiple forms simultaneously. Bypassing the need to specify a sequence of linking steps, it potentially reduces the risk of cumulative error and therefore may be more robust to scale drift (Haberman, 2010).

Simultaneous linking involves linking multiple test forms together by sharing common items, although not necessarily the same common items across all administrations. Item parameter estimates obtained through separate calibration must be available at the time of simultaneous linking. The process involves an iterative procedure, which makes the computational difficulty of simultaneous linking more manageable.

Haberman (2009) presented a description of the two parameter-logistic (2-PL) model for simultaneous linking. The scale transformation coefficients are determined through a two-step procedure. Suppose that there are T administrations and t represents the t^{th} administration such that $t = 1, \ldots, T$. In the first step, the slope (A_t) is found by minimizing the following equation:

$$\sum_{t=1}^{T} \sum_{j \in J_t} [\log \hat{a}_{jt} - \log \hat{A}_t - \log \hat{a}_j]^2, \tag{1}$$

where $\hat{a}_j$ represents the estimated item discrimination parameter for item j on the base scale, and $\hat{a}_{jt}$ represents the estimated item discrimination parameter for item j in administration t.

In the second step, the intercept (B_t) is found using the previously determined slope A_t by minimizing the following equation:

$$\sum_{t=1}^{T} \sum_{j \in J_t} [\hat{b}_{jt} \hat{A}_t + \hat{B}_t - \hat{b}_j]^2, \tag{2}$$

where $\hat{b}_j$ represents the estimated item difficulty parameter for item j on the base scale, and $\hat{b}_{jt}$ represents the estimated item difficulty parameter for item j in administration t.

After obtaining the scale transformation coefficients, the new-form parameter estimates can be placed on the base scale using the following equations:

$$\hat{a}'_{jt} = A_t \hat{a}_j \tag{3}$$

$$\hat{b}'_{jt} = (\hat{b}_j - B_t)/A_t \tag{4}$$

1.2 Stocking-Lord Scale Transformation

Stocking and Lord (1983) criticized traditional moment methods, such as mean/mean and mean/sigma, for not incorporating all item parameter estimates simultaneously. For instance, the mean/sigma method calculates scale transformation coefficients using only item difficulty estimates. These coefficients are then used to transform item discrimination estimates and abilities. This process can be problematic as two significantly different item difficulty estimates, in conjunction with other item and person estimates, might yield very similar item characteristic curves (Kolen & Brennan, 2014). In contrast, the Stocking-Lord method estimates transformation coefficients using test characteristic curves. By using all item parameter estimates simultaneously, this method generally produces more accurate results than moment methods (Hanson & Béguin, 2002; Kolen & Brennan, 2014).

The Stocking-Lord method and Haebara method are both characteristic curve methods, with minor differences. The core principle behind these methods is that each item calibration will produce an estimated item characteristic curve. Assuming no error, a linear transformation should make the curves from two different scales align (Stocking & Lord, 1983). In other words, the probability of an examinee answering an item correctly should remain constant, regardless of the scale used. Characteristic curve methods aim to minimize the difference between the item characteristic curves to find the optimal scale transformation coefficients (Kolen & Brennan, 2014).

According to the Stocking-Lord method, in a 2-PL IRT model, the square difference of sums over items, for a given θ_i, is expressed in the following equation:

$$SLdiff(\theta_i) = \left[\sum_{j:V} p_{ij}(\theta_{Ji}; \hat{a}_{Jj}, \hat{b}_{Jj}) - \sum_{j:V} p_{ij}(\theta_{Ji}; \frac{\hat{a}_{Ij}}{A}, A\hat{b}_{Ij} + B) \right]^2, \tag{5}$$

Here, I and J represent old and new scales, and V is the number of common items. A and B represent the slope and intercept of scale transformation. $\hat{a}_{Jj}$ and $\hat{b}_{Jj}$ represent the item discrimination and difficulty estimates for item j on scale J, while $\hat{a}_{Ij}$ and $\hat{b}_{Ij}$ represent the item discrimination and difficulty estimates for item j on scale I.

$SLdiff$ is then cumulated over all examinees. A combination of A and B can be found to minimize the given equation:

$$SLcrit = \sum_i SLdiff(\theta_i), \tag{6}$$

Details on computation can be found in the original paper Stocking and Lord (1983).

1.3 Motivation and Research Objectives

The Stocking-Lord (SL) method and simultaneous linking (SM) share similarities. Both require the estimates for each test form to be linked through separate calibration, and both involve an iterative approach in the minimization process. However, unlike SL, SM is considered a generalized log-mean/mean method which faces the same criticism for not incorporating all parameter estimates concurrently. Specifically, SM uses the item discrimination estimates to find the scale transformation slope and uses this slope to find the intercept. On the other hand, the SL transformation was originally designed for two test forms, whereas SM can link multiple forms simultaneously to reduce the error accumulation resulting from a long linking chain. The comparative performances of these two methods are yet to be investigated.

This study endeavors to enhance the current understanding of how simultaneous linking performs under diverse scenarios. Furthermore, this study underscores error accumulation across administrations and its effect on item parameter recovery. To address these research objectives, a simulation study was conducted. Specifically, the study aims to address two research questions. First, how does SM compare to SL method in terms of their performance under different IRT models, linking designs, and ability distributions? Second, does the repeated use of the same items introduce more bias in the item parameter estimates?

1.4 Previous Studies

Previous research has extensively compared performances of various linking methods including separate, concurrent, and fixed calibration methods. Kim and Cohen (1998) found that separate calibration was preferred over concurrent calibration when the number of common items was small, whereas the two methods performed similarly when the number of common items was large. Lee and Ban (2009) compared various IRT linking procedures in a random groups design and found that separate calibration, in general, outperformed concurrent calibration. Lee and Ban (2009) also found that, among separate calibration procedures, the Haebara method produced lower linking error than the Stocking-Lord method, whereas concurrent calibration outperformed separate calibration when all the samples were from the same population following a standard normal distribution. However, no scaling might be a better alternative in this scenario.

Kang and Petersen (2012) emphasized the significance of a proper implementation of fixed calibration, a procedure that involves fixing common item parameters at the values obtained from the previous calibration, and then freely estimating parameters of unique items in the new form. Kang and Petersen found that when fixed calibration was correctly implemented via using multiple EM cycles and updating the prior ability distribution multiple times during calibration, the results were comparable to those for the separate and concurrent calibration methods. However, an incorrect implementation of fixed calibration may introduce severe bias to the results.

Previous research has provided valuable insights into the performance of different IRT linking methods. However, there is a lack of understanding regarding the performance of simultaneous linking in comparison to these methods. Among the limited studies that have investigated the performance of simultaneous linking, Robitzsch (2020a) highlighted the similarity between the simultaneous linking and alignment methods. Proposed by Muthén and Asparouhov (2014), the alignment method estimates group-specific factor means and variances without requiring full measurement invariance, allowing for multiple group comparisons under confirmatory factor analysis (CFA) and IRT models. As linking functions, the underlying principle of both simultaneous and alignment methods is to minimize the deviation between group-specific item parameters. Based on the study, Robitzsch (2020a) concluded that the two methods performed similarly in estimating group means, with robust simultaneous linking having a slight edge over the alignment method. Unfortunately, this study did not examine recovery of item parameters.

Simultaneous linking has demonstrated promising results when compared to separate calibration in conjunction with Stocking-Lord transformation. Lu and Antal (2022) acknowledged the convenience and efficiency of simultaneous linking, in addition to its advantages in protecting against scale drift. Specifically, simultaneous linking outperformed the Stocking-Lord procedure when items were administered more than three times (preferably more than five times) and in the presence of nonsystematic parameter drift. Furthermore, simultaneous linking demonstrated maximum advantages when forms were administered in a chained design. Considering the substantial influence of data collection procedures on linking outcomes, the present study aims to compare effectiveness of the simultaneous linking and Stocking-Lord methods under different data collection designs.

2 Method

2.1 Study Factors and Conditions

Data simulation and analyses were conducted in R (R Core Team, 2023). The study considered the following factors: (1) IRT models used to generate and calibrate data, either 1-PL or 2-PL; (2) examinees' ability distributions; (3) the linking design, either chained or 5-replication; and (4) the percentage of common items, either 40% for the chained design or 60% for the 5-replication design. The study considered a total of 15 administrations ($T = 1, \ldots, 15$). Throughout the study, sample size and test length were fixed to 1,500 and 50 items, respectively, for each administration. These conditions were kept fixed based on the hypothesis that altering them would not significantly impact the overall comparative patterns between the two linking methods. For examinees' ability distributions, the study considered the following conditions: (a) $N(0,1)$ for all 15 administrations, (b) $N(\mu,1)$ with $\mu = 0$ to .7 with increments of .05 across 15 administrations, and (c) $N(\mu,1)$ with $\mu = 0$ to -.7 with increments of -.05 across 15 administrations. For each crossed study condition, we generated 1,000 data sets ($R = 1{,}000$).

To enable the comparison of common and unique items, parallel test forms were created to be as similar as possible to eliminate confounding effects due to form differences. Item blocks, each with 10 items, were constructed so that they had similar means and standard deviations of item discrimination and item difficulty parameters. The total number of blocks required depended on the linking design.

Figure 1 presents the chained design that consists of 47 blocks with 40% common items between two adjacent administrations. The year 1 population associated with the $N(0,1)$ ability distribution serves as the base scale. When SL is used for scale transformation, a chain is formed such that item parameter estimates for later administrations go through multiple links to be linked to the scale of year 1. By contrast, SM requires only one run. For instance, to link year 15 to year 1 using the SL method, a sequential linking process occurs where year 15 is linked to year 14, which in turn is linked to year 13, and so forth, until year 2 is linked to year 1.

Figure 2 presents a linking design where each of the first seven blocks is administered five times. This design is, thus, referred to as the 5-replication design, hereafter. In order to assess the accuracy of item parameter recovery in the context of repeated use of items, this particular design comprises a total of 33 blocks, corresponding to 15 administrations. While our primary attention is directed towards the initial seven blocks, errors for the remaining blocks are also provided for a comprehensive evaluation. A total of 33 blocks are required to secure enough items that are utilized five times across 15 administrations while introducing two new blocks for each administration.

When SL transformation is used for the 5-replication design, multiple linking solutions can be considered. For example, in the case of linking year 4 to year 1, two different approaches can be taken. The first approach uses block 1 to link year 4 directly to year 1. The second approach involves a two-step linking process: step 1) link year 4 to year 2 using blocks 1 and 7, and step 2) link year 2 to year 1 using blocks 1, 3, and 5. For each year/administration, among all possible approaches, the approach that required the smallest number of linking steps (i.e., the shortest link) was consistently selected as the preferred linkage.

The 5-replication design has 60% common items between two adjacent administrations; however, the total percentage of common items involved in scale linking differed depending on the target year, ranging from 20% to 60%. For example, to link year 15 back to year 1, block 4 was utilized as the common block, corresponding to having 20% common items. On the other hand, linking year 2 to year 1 involved blocks 1, 3, and 5 which correspond to having 60% common items. It is noteworthy that errors in the 5-replication design are not expected to accumulate, as it has only one linkage regardless of the target year. In contrast, in the chained design, errors are expected to accumulate due to multiple linkages. Errors from 60% common items are anticipated to exhibit a similar pattern to those in 40%, albeit of a smaller magnitude. Only 40% common items for the chained design or 60% common items for the 5-replication design were considered, as the percentage (whether 40% or 60%) is not expected to change the comparative performances of the two linking methods.

2.2 Data and Calibration

In this study, realistic item parameters were obtained by calibrating a real dataset using the 2-PL model. These parameters were then used to construct parallel blocks and simulate data. Both calibration and data generation were conducted using the *irtoys* package (Partchev, Maris, & Hattori, 2022). The scaling constant was set to 1 for all IRT models. Separate calibration was performed, and different administrations were linked through common items using the *plink* package (Weeks, 2010). Simultaneous linking, on the other hand, was conducted using the *sirt* package (Robitzsch, 2022).

For the 1-PL condition, the same dataset was calibrated using the 1-PL model, with item discrimination parameters fixed at 0.57, which was one realistic value of the base scale. Parallel blocks were then constructed using the acquired item difficulty parameters. Subsequently, for each of the study conditions, responses for the 15 administrations were simulated using the item parameters associated with blocks, and the simulated data were calibrated using the 1-PL model. For each administration, item discrimination parameters were also estimated under the restriction that they were the same for all items administered at the same time. However, the focus of this condition was solely on the recovery of item difficulty parameters. The linking procedures were identical to those of the 2-PL condition. The means and standard deviations of item parameters for each condition are summarized in Table 1.

3 Evaluation Criteria

3.1 Evaluation Approaches

The study employed two evaluation approaches to compare results between the two linking methods. The first approach involved organizing the results by block. The approach evaluated final item parameter estimates for each block upon the end of year 15. The second approach involved organizing the results by year. This approach facilitates the understanding of result patterns over time including the accumulation of errors.

It is important to mention that the SM method yields a single set of estimates, whereas the SL method produces a set of estimates for each time an item is used in the linking process. This means that at the end of linking all administrations to the base scale of year 1, each common item is associated with multiple sets of item parameter estimates in the SL method. To facilitate a fair comparison between the two methods, the item parameter estimates produced by SL were averaged and used for the final evaluation. For instance, based on chained design (Figure 1), when year 2 was linked to year 1, there were two sets of item parameter estimates for blocks 4 and 5: the first set consists of item parameter estimates from year 1 and the second set consists of item parameter estimates from year 2 that are scale-transformed to the scale of year 1. Thus, for the final evaluation, the study used the average of the two estimates.

Due to the nature of the chained design, all common blocks were used twice. When results were organized by block, item parameter estimates produced by SL were averaged across the two usages for each common block and compared against item parameters. In contrast, the SM method yields a single set of item parameter estimates for each item; there is no need to average item parameter estimates across the two usages. Results organized by year present errors in item parameter estimates for five blocks associated with each administration. For each year, errors were averaged over five blocks associated with the administration upon its administration.

Similar to the chained design, results for the 5-replication design were organized either by block or by year. For results organized by block, final item parameter estimates for the SL method were obtained by averaging the estimates obtained from all administrations. For example, the final item parameter estimates for block 1 were the averages of estimates from the year 1, year 2, year 4, year 8, and year 12 administrations. In contrast, with the SM method, linking was conducted simultaneously across all 15 administrations, yielding a single set of final item parameter estimates. On the other hand, results organized by year were obtained by using data accumulated up to a specific targeted year, resulting in 15 sets of results representing 15 administrations for both linking methods (Figure 2).

3.2 Accuracy in Item Parameter Estimates

Item parameter estimates were evaluated in terms of squared bias (SB), variance, mean squared error (MSE). SB is defined as the squared value of the difference between estimated and true parameters; and variance captures the deviation from the mean of estimated parameters. MSE is the sum of SB and variance. SB, variance, and MSE are given by:

$$\mathrm{SB}_{pi} = \frac{\sum_{r=1}^{R}(\hat{p}_{ri} - p_i)^2}{R}, \tag{7}$$

$$\mathrm{Variance}_{pi} = \frac{\sum_{r=1}^{R}(\hat{p}_{ri} - \bar{\hat{p}}_i)^2}{R}, \tag{8}$$

$$\mathrm{MSE}_{pi} = \mathrm{SB}_{pi} + \mathrm{Variance}_{pi}, \tag{9}$$

where p refers to item parameters, either discrimination or difficulty; p_i and $\hat{p}_{ri}$ refer to item parameter and parameter estimate for item i from the r^{th} replication, respectively; and $\bar{\hat{p}}_i$ refers to the average of R estimates for item i.

3.3 Recovery of Test Characteristic Curves

Furthermore, test characteristic curves (TCC) based on item parameter estimates were evaluated in a similar manner. Overall SB, variance, and MSE of TCC estimates are given by:

$$\mathrm{SB}_{TCC} = \sum_{q=1}^{Q} \mathrm{SB}_{TCC}(\theta_q) \times w(\theta_q), \tag{10}$$

$$\text{Variance}_{TCC} = \sum_{q=1}^{Q} \text{Variance}_{TCC}(\theta_q) \times w(\theta_q), and \tag{11}$$

$$\text{MSE}_{TCC} = \sum_{q=1}^{Q} \text{MSE}_{TCC}(\theta_q) \times w(\theta_q), \tag{12}$$

where conditional SB, variance, and MSE in TCC can be obtained as follows:

$$\text{SB}_{TCC}(\theta_q) = \left(\frac{\sum_{r=1}^{R} \widehat{\text{TCC}}_r(\theta_q)}{R} - \text{TCC}(\theta_q) \right)^2 = \left(\overline{\widehat{\text{TCC}}}(\theta_q) - \text{TCC}(\theta_q) \right)^2, \tag{13}$$

$$\text{Variance}_{TCC}(\theta_q) = \frac{\sum_{r=1}^{R} (\widehat{\text{TCC}}_r(\theta_q) - \overline{\widehat{\text{TCC}}}(\theta_q))^2}{R}, and \tag{14}$$

$$\text{MSE}_{TCC}(\theta_q) = \text{SB}_{TCC}(\theta_q) + \text{Variance}_{TCC}(\theta_q). \tag{15}$$

In the above equations, $\theta_q (q = 1, \ldots, Q)$ is the q^{th} ability point; the weight associated with the $N(0,1)$ density function is denoted by $w\,(\theta_q)$. TCC (θ_q) and $\widehat{\text{TCC}}_r\,(\theta_q)$ refer to the TCC based on item parameters at θ_q and the TCC based on item parameter estimates from the r^{th} replication at θ_q, respectively; and $\overline{\widehat{\text{TCC}}}\,(\theta_q)$ refers to the average of R TCC estimates at θ_q. $\widehat{\text{TCC}}_r\,(\theta_q)$ and $\overline{\widehat{\text{TCC}}}\,(\theta_q)$ are given by:

$$\widehat{\text{TCC}}_r(\theta_q) = \sum_{i=1}^{n} P(Correct\ Response|\theta_q, \hat{a}_{ri}, \hat{b}_{ri}), \tag{16}$$

$$\overline{\widehat{\text{TCC}}}(\theta_q) = \frac{\sum_{r=1}^{R} \widehat{\text{TCC}}_r(\theta_q)}{R}, \tag{17}$$

where n refers to the number of items; and $\hat{a}_{ri}$ and $\hat{b}_{ri}$ refer to the discrimination and difficulty parameter estimates for item i from the r^{th} replication, respectively. As mentioned earlier, the number of replications was $R = 1,000$. θ_q is a sequence of numbers representing ability ranging from -4 to 4 with an increment of .01.

4 Results

Since different conditions for ability distribution did not alter the overall comparative performance patten of the two linking methods, this section presents results for only one ability distribution condition. For the 2-PL model condition, results for the increasing ability condition are reported, and results for the constant ability condition are presented for the 1-PL model condition. However, it should be noted that varying conditions of ability distribution did have an impact on the magnitude of errors.

4.1 2-PL Model

4.1.1 Chained Design (Increasing Ability Condition)

Figure 3 presents SB of item discrimination parameters by block for the increasing ability condition when the 2-PL model was used for calibration under the chained design. The subfigures were organized by all blocks, common blocks only, and unique blocks only. Note that higher block numbers are associated with later administrations. Based on Figure 3, SM consistently yielded smaller SB than SL, indicating better accuracy in recovering item discrimination parameters. The differences between the two linking methods became more evident at later blocks in comparison to earlier blocks.

Similarly, Figure 4 presents SB for item difficulty parameters by block for the increasing ability condition. Notably, the overall pattern flipped compared to the previous observations. For difficulty parameters, SBs for the SL method were smaller than those for the SM method for all blocks. The difference between the two methods became more pronounced for blocks administered later; in the case of SL method, these blocks required more chains to link back to the scale of year 1. However, it is worth noting that the differences in SB between the two linking methods were extremely small in magnitude.

Figures 5 to 8 display the variance and MSE for item discrimination and difficulty parameter estimates, respectively, for the increasing ability condition. The SL method yielded slightly smaller variances and MSEs than the SM method did. However, for all blocks, differences in variances and MSEs between the SM and SL methods seemed to be very small for both item parameters.

The next set of results were structured by utilizing all the blocks in their respective years. In each plot, the horizontal axis represents year; for each year, errors were averaged over five blocks associated with the administration. To enable a fair comparison of the two linking methods, 15 sets of estimates were obtained using the SM and SL methods. The sets included data from additional years in the linking procedure, with the first set comprising estimates obtained from calibration for year 1. The second set included data from years 1 and 2, the third set included data from years 1, 2, and 3, and so on.

Figures 9 and 10 summarize the results for the increasing ability condition, specifically in terms of SB, variance, and MSE for item discrimination and item difficulty parameters. The results suggest that SM may have an advantage in recovering item discrimination parameters but may struggle with item difficulty parameter recovery in this condition when compared to SL. SM demonstrated lower SB for item discrimination and higher SB for item difficulty in comparison to SL. Additionally, SM exhibited higher variance and MSE for both item parameters.

According to the TCC criteria, SL outperformed SM slightly. The TCC results were organized by block and by year. Based on Figure 11, for the blocks used in later administrations, SM exhibited larger SB, which was more evident in the unique blocks. However, the pattern fluctuated for the common blocks. Similarly, when the results were organized by year, Figure 12 showed that SBs for the SM method were larger than those for the SL method for year 6 and later administrations. Variances and MSEs for SM were also larger than those for SL across all administrations. As suggested by the increasing pattern of lines in all figures, both the SM and SL methods exhibited error accumulation over time for the chained design.

4.1.2 5-Replication Design (Increasing Ability Distribution)

Figures 13 and 14 present SB of item discrimination and difficulty parameter estimates organized by block for the increasing ability distribution. The differences in SB between the two linking methods seemed extremely small. However, consistent with the chained design, the SM method tended to produce lower bias in recovering item discrimination parameters, regardless of whether the block was common or unique. Meanwhile, for both common and unique blocks, SBs in item difficulty parameter estimates for the SL method were lower than those for the SM method.

Figures 15 to 18 present variance and MSE of item discrimination and difficulty parameter estimates, respectively, for the increasing ability condition. Based on the figures, regardless of item parameters, variances and MSEs for all blocks were similar for the SM and SL methods. However, in contrast to the results for the chained design, variances and MSEs for the SM method under the 5-replication design were marginally lower than those for the SL method for both item parameters.

Figure 19 presents SB, variance, and MSE in item discrimination parameter estimates organized by year for the increasing ability condition. Compared to the results for the chained design, Figure 19 does not exhibit evident increasing patterns in errors over time. These results validate the earlier statement that errors do not accumulate in the 5-replication design because the design involves only one linkage for both the SM and SL methods. The results suggest that the SM method performed better than the SL method in terms of SB, variance, and MSE in item discrimination parameter estimates.

Figure 20 presents SB, variance, and MSE in item difficulty parameter estimates organized by year for the increasing ability condition. Based on Figure 20, SBs tended to be similar for the SM and SL methods; SBs for the SM method were slightly higher at year 4, year 12, and year 13, but lower for the remaining administrations than those for the SL method. Conversely, SM consistently produced lower variance and MSE for all 15 administrations than SL.

Figure 21 presents SB in TCC by block for the increasing ability condition. Based on the results, the recovery of item difficulty parameters seemed to heavily influence the recovery of TCC. For all blocks, SB in TCC for SM tended to be slightly larger than those for SL. Note that SBs in difficulty parameter estimates were also slightly larger for SM than for SL (see Figure 14). When the results were organized by year, Figure 22 shows that SM resulted in lower SB in TCC for the majority of administrations, with a few exceptions, which is similar to the pattern in SB of item difficulty parameter estimates as shown in Figure 20. Additionally, at year 4 and later administrations, SM exhibited lower variance and MSE than SL did.

Figures 23 and 24 present SB of item discrimination and difficulty parameter estimates for the first seven blocks that were repeatedly administered five times. The horizontal axis denotes the number of replications, while the vertical axis represents the magnitude of SB, with each colored line representing a block. The results suggest that both linking methods exhibited similar performances, with an observable decreasing trend as the blocks were used repeatedly. It should be noted that the blocks themselves are not comparable since each block was used in different administrations. Importantly, in the absence of parameter drift, recurring use of the same blocks over time did not lead to the accumulation of error in parameter estimates. In fact, the more the same blocks were used, the more accurate the results became. Moreover, although not presented here, the variance and MSE also decreased as the blocks were repeatedly used for both linking methods. Overall, these findings suggest that using the same blocks repeatedly is an effective strategy for improving the accuracy of parameter recovery in linking.

It should also be noted that while the overall trend of the results remains consistent for all ability distributions, the magnitude of SB differences between two linking methods were more evident in situations where the ability distributions were constant or decreasing, as opposed to increasing. Additionally, the use of the SM method resulted in smaller variance and MSE in most conditions.

4.2 1-PL Model

4.2.1 Chained Design (Constant Ability Distribution)

The linking and evaluation procedures for the 1-PL model are analogous to those for the 2-PL model. However, for the 1-PL model, we are solely interested in item difficulty parameters. As previously mentioned, varying ability distribution conditions did not significantly alter the overall patterns of results. Thus, this section presents results for the constant ability distribution to highlight differences between the two linking methods using the 1-PL model. All of the results obtained under the chained design for the 1-PL model are consistent with those for the 2-PL model.

In Figure 25, SB of item difficulty parameter estimates is organized by block for the constant ability condition. The results show that, for the majority of blocks (especially block 12 and later blocks), the SM method yielded larger SB in comparison to the SL method, regardless of whether the blocks were common or unique. However, based on Figures 26 and 27, SM exhibited smaller variance and smaller MSE for all blocks than SL did, although the differences were comparatively small. When results were organized by year, the SBs produced by both methods were very similar in magnitude. However, SM demonstrated smaller variance and MSE compared to SL (Figure 28).

Furthermore, the study examined the recovery of TCC by block. Based on Figure 29, SBs for SM tended to be slightly larger than those for SL. This result is expected, as the SL method showed slightly better recovery in difficulty parameters compared to the SM method. When the TCC results were organized by year, Figure 30 shows that the performances of both linking methods were similar, with SM producing slightly larger SBs and smaller variances and MSEs than SL.

4.2.2 5-Replication Design (Constant Ability Distribution)

This section compares the performance of the SM and SL methods for the 1-PL model under the 5-replication design with the constant ability distribution condition. According to Figure 31, neither linking method consistently outperformed the other in terms of SB. In half of the blocks, SM yielded smaller SBs than SL, while SL produced smaller SBs for the other half. This pattern was observed for both common and unique blocks. In contrast, based on Figures 32 and 33, variances and MSEs of difficulty parameter estimates were consistent with the previous observations; namely, the errors for SM tended to be smaller the those for SL for all blocks.

When examining the results organized by year, Figure 34 shows that the SM method yielded smaller SBs in most years, except for years 4 and 8. Furthermore, the SM method yielded smaller variances and MSEs compared to the SL method, particularly when the results were organized by year rather than by block. In addition, the errors seem relatively consistent over the years without showing an evident increasing pattern. This again suggests that there was no accumulation of error over time for the 5-replication design.

The results of the TCC criterion were heavily influenced by the recovery of item difficulty parameters. Figures 35 and 36 present results organized by block and by year, respectively. When the results were organized by block, the SM method yielded smaller values of SB for half of the blocks, irrespective of whether they were common or unique. When the results were organized by year, SBs in TCC for the SM method were smaller for the majority of years, with the exception of years 4 and 8, than those for the SL method. Moreover, SM consistently exhibited smaller variances and MSEs for TCC in comparison to SL.

Figures 37 presents SB of difficulty parameter estimates for the first seven blocks that were repeatedly administered five times. The results indicate that item recovery accuracy improves with the repeated usage of the same blocks, in the absence of parameter drift. Both linking methods exhibit a decreasing trend in SB, with the SL method demonstrating a slight advantage under this criterion. This advantage is evident as the majority of lines appeared lower in position compared to SM, indicating better performance in terms of SB for the SL method.

4.3 Relationship between Item Parameters and Accuracy of Parameter Recovery

This section examines the relationship between the accuracy of item parameter recovery and the magnitude of item parameters. Specifically, Figure 38 illustrates the correlation between item discrimination parameters and SB of item discrimination parameter estimates. The plot is based on the 2-PL chained design with the decreasing ability distribution condition. Again, varying ability distribution did not affect the comparative performances of the two linking methods. Items from all 47 blocks were ranked based on their item discrimination parameters. The lowest

30% were categorized as “Low,” the middle 40% as “Medium,” and the highest 30% as “High.” The colored dots on the plot represent this ranking, with the darkest blue dots indicating items in the “High” category. The majority of the darkest blue dots are associated with relatively larger SBs, indicating that there exists a positive correlation between the item discrimination parameters and the SB of the item discrimination estimates. The correlation coefficient was 0.34 for the SM method and 0.39 for the SL method.

Similar results were observed for item difficulty parameters (see Figure 39). The correlation between item difficulty parameters and SB of item difficulty estimates was .36 for SM and .46 for SL, indicating a relatively strong positive relationship. Even stronger correlations were observed in terms of variance and MSE. Figures 40 and 41 present relationships between parameter values and variances in estimates for discrimination and difficulty parameters, respectively. Based on Figure 40, the correlation between item discrimination parameters and the variance of the estimates was .67 for both SM and SL. Figure 41 shows that the correlation between item difficulty parameters and the variance of their estimates was .50 for both SM and SL. Furthermore, Figures 42 and 43 present relationships between parameter values and MSEs in estimates for discrimination and difficulty parameters, respectively. The correlation was .67 and .66 for SM and SL, respectively. Finally, the correlation between item difficulty parameters and the MSE of the estimates was .51 for SM and .50 for SL.

The aforementioned findings pose a question regarding the balance between the quality of test items and potential linking errors. Although highly discriminating items (and items with increased difficulty in certain situations) are generally preferred, the utilization of such items may lead to notable linking errors. The determination of the right balance between these factors remains an area requiring further investigation.

5 Discussion

The primary goal of this study is to explore how simultaneous linking compares to Stocking-Lord method under the same linking conditions. Additionally, we aim to investigate whether the repeated use of the same blocks introduces bias into item parameter estimates. Our key findings are summarized as follows:

- For the 2-PL model, SM produced lower SB values for item discrimination parameters and higher SB values for item difficulty parameters compared to SL for both chained and 5-replication designs.
- For the 1-PL model, SM produced higher SB values for item difficulty parameters than SL in the chained design, but no clear pattern was found in the 5-replication design.
- SM produced smaller variance and MSE values for nearly all conditions, with the exception being the 2-PL chained design.
- The SB values of the TCC criteria fluctuated, yet heavily driven by the item difficulty parameter.
- Varying ability distributions did impact the magnitude of bias and other evaluation criteria, but did not change the overall performance pattern of the two linking methods.
- Under the assumption of no parameter drift, the repeated use of the same items did not introduce bias to the item parameters. Indeed, the more you use the items, the more accurate the results can be.
- The differences in performance between SM and SL were minimal.

This study's major finding is that SM exhibited better performance in estimating item discrimination parameters, while SL was better at recovering item difficulty parameters under the 2-PL model. Notably, SM requires a two-step minimization process to estimate the slope and intercept for scale transformation separately. Initially, the slope is estimated first, and subsequently it is used to find the intercept. In contrast, SL estimates both slope and intercept concurrently. This procedural distinction could potentially explain the differential performance in recovering the discrimination and difficulty parameters between the two methods. However, further research would be needed to delve deeper into this phenomenon.

Both SM and SL demonstrated a gradual accumulation of errors over time in the chained design. Although SM showed certain advantages in specific scenarios, the overall amount of error accumulated was comparable between the two linking designs, and the magnitude of error was exceedingly small. It is important to acknowledge that claiming SM as a method that completely prevents error accumulation may be problematic, as it solely relies on a single model specification. In reality, both SM and SL have the potential to mitigate the risk of substantial cumulative errors. Additionally, we found that utilizing the same blocks repeatedly proved to be an effective strategy for enhancing the accuracy of parameter recovery in linking, as evidenced by the observed decreasing trend in error across repetitions. This approach may also allow for a reduction in the number of items required in the item bank. However, it is crucial to consider the potential security implications of overexposing test items. Furthermore, we face the challenge of determining whether to update or retain parameter estimates for recurring items.

Parameter drift was intentionally omitted from the scope of the current simulation study due to the inherent challenge associated with establishing meaningful evaluation criteria when parameter drift is present. Variability and shifts in item parameters over time make it difficult to establish a consistent criterion for evaluating the performance of different linking methods. In future research, it would be worthwhile to develop effective criteria for evaluating the impact of parameter drift on linking through simulation studies.

Considering the ease of catching parameter drift, the use of separate calibration with SL transformation can be a favorable option. By evaluating item parameter estimates at each administration, any abnormal deviation can be identified. From an operational standpoint, the SM method offers the advantage of requiring only one estimation process. However, it should be noted that this approach necessitates having access to all item parameter estimates from all administrations at the time of linking, which may not always be feasible. In such circumstances, separate calibration procedures become more viable alternatives.

As discussed earlier, the SL method produces multiple sets of estimates for the common items, making it difficult to compare directly to SM. Considering the nature of the SM method, it may be more suitable to compare it with linking methods specifically designed for multiple groups. Researchers have conducted investigations into the expansion of widely employed linking techniques through separate calibrations for multiple groups or have introduced novel methodologies to address the challenges posed by multiple test forms. For instance, Battauz (2017) proposed a generalization of the mean-geometric mean, mean-mean, Haebara, and Stocking-Lord methods to multiple test forms. Robitzsch (2020b) introduced the robust Haebara linking approach, which incorporates a power loss function to compare many groups. Furthermore, Battauz and Leôncio (2023) developed a likelihood-based approach to account for the heteroscedasticity and correlation of the item parameter estimates of multiple test forms. While these approaches are not within the scope of our current investigation, future research would be needed to assess and evaluate the effectiveness of these methods.

The items used in the present study were dichotomous, and the next logical step would be to extend the SM method to polytomous items and mixed format tests. Additionally, while our study scenarios focused solely on horizontal scaling, it would be intriguing to explore how SM performs under vertical scaling scenarios. This expanded exploration will contribute to a more comprehensive understanding of both the strengths and limitations of SM compared to SL across a variety of linking scenarios.

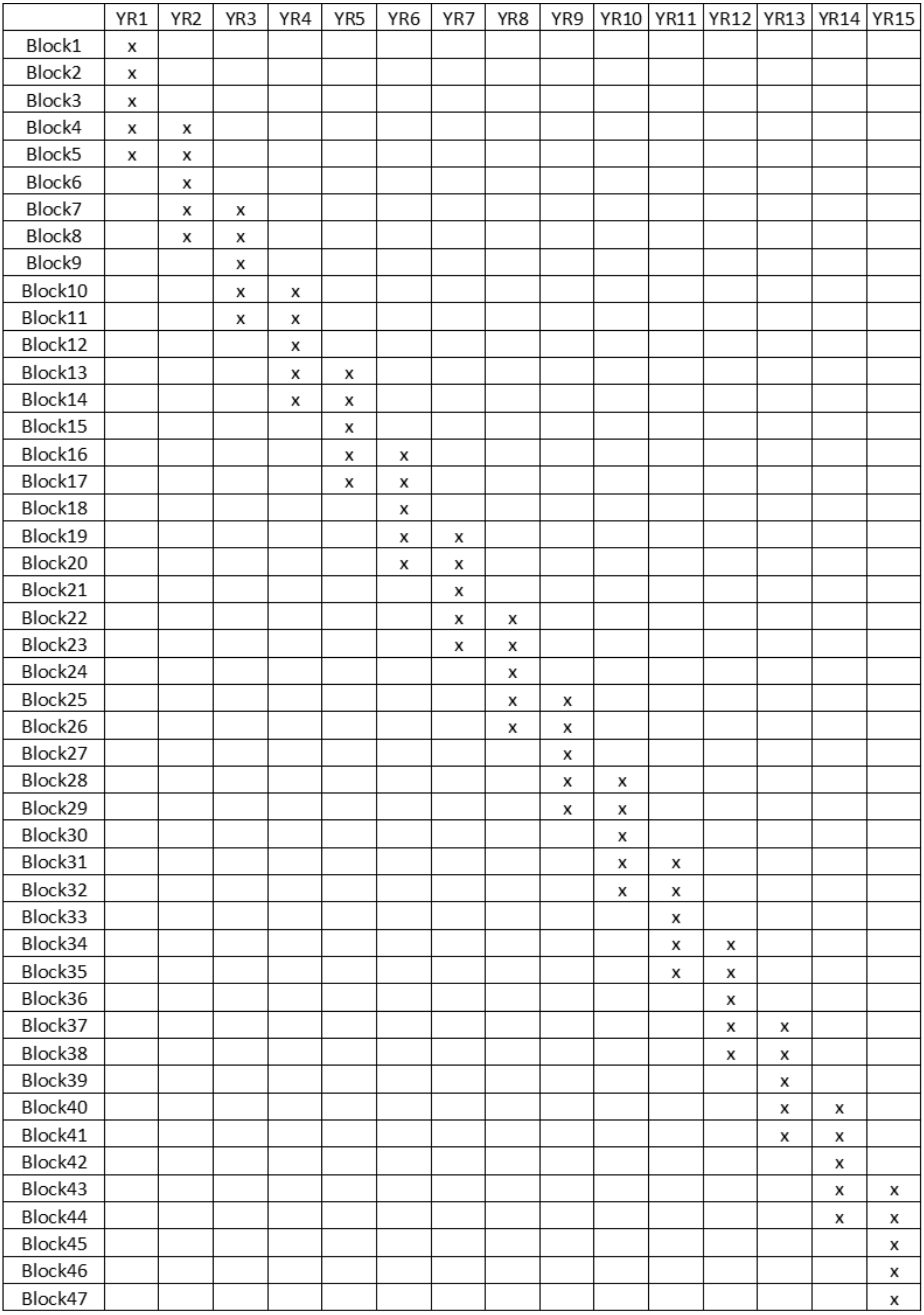
This study has enhanced our understanding of the relative performance of SM when compared to separate calibration through SL transformation. While certain patterns of differential performance were noted, overall, both methods demonstrated comparable performance. Our findings highlight the importance of thoughtful consideration of various factors, such as the IRT model, item banking, and linkage design, when choosing a linking method.

Table 1
Average Block Descriptive Statistics of Item Parameters in Each Linking Design

	2-PL				1-PL			
	Chained		5-Replication		Chained		5-Replication	
	a	*b*	*a*	*b*	*a*	*b*	*a*	*b*
Mean	.914	.439	.914	.439	.570	-.004	.570	-.006
SD	.231	.749	.231	.749	.000	1.193	.000	1.194
Maximum	.924	.449	.924	.449	.570	.029	.570	.026
Minimum	.906	.430	.906	.430	.570	-.028	.570	-.028

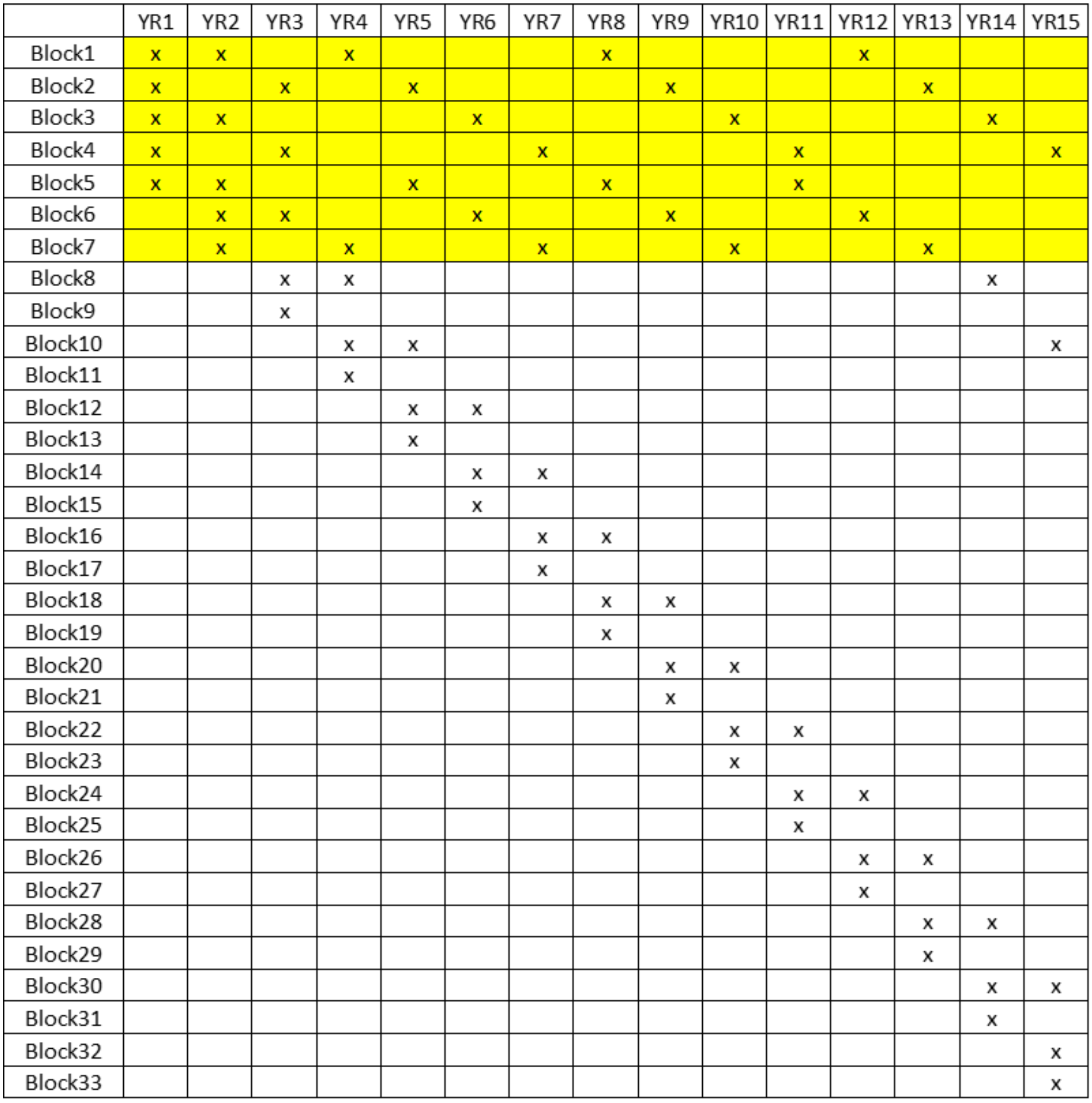
Note. Notations *a* and *b* refer to discrimination and difficulty parameters, respectively.

Figure 1
Chained Design



	YR1	YR2	YR3	YR4	YR5	YR6	YR7	YR8	YR9	YR10	YR11	YR12	YR13	YR14	YR15
Block1	x														
Block2	x														
Block3	x														
Block4	x	x													
Block5	x	x													
Block6		x													
Block7		x	x												
Block8		x	x												
Block9			x												
Block10			x	x											
Block11			x	x											
Block12				x											
Block13				x	x										
Block14				x	x										
Block15					x										
Block16					x	x									
Block17					x	x									
Block18						x									
Block19						x	x								
Block20						x	x								
Block21							x								
Block22							x	x							
Block23							x	x							
Block24								x							
Block25								x	x						
Block26								x	x						
Block27									x						
Block28									x	x					
Block29									x	x					
Block30										x					
Block31										x	x				
Block32										x	x				
Block33											x				
Block34											x	x			
Block35											x	x			
Block36												x			
Block37												x	x		
Block38												x	x		
Block39													x		
Block40													x	x	
Block41													x	x	
Block42														x	
Block43														x	x
Block44														x	x
Block45															x
Block46															x
Block47															x

Figure 2
5-Replication Design



	YR1	YR2	YR3	YR4	YR5	YR6	YR7	YR8	YR9	YR10	YR11	YR12	YR13	YR14	YR15
Block1	x	x		x				x				x			
Block2	x		x		x				x				x		
Block3	x	x				x				x				x	
Block4	x		x				x				x				x
Block5	x	x			x			x			x				
Block6		x	x			x			x			x			
Block7		x		x			x			x			x		
Block8			x	x										x	
Block9			x												
Block10				x	x										x
Block11				x											
Block12					x	x									
Block13					x										
Block14						x	x								
Block15						x									
Block16							x	x							
Block17							x								
Block18								x	x						
Block19								x							
Block20									x	x					
Block21									x						
Block22										x	x				
Block23										x					
Block24											x	x			
Block25											x				
Block26												x	x		
Block27												x			
Block28													x	x	
Block29													x		
Block30														x	x
Block31														x	
Block32															x
Block33															x

Figure 3

SB of Item Discrimination by Block (2-PL, Chained Design)

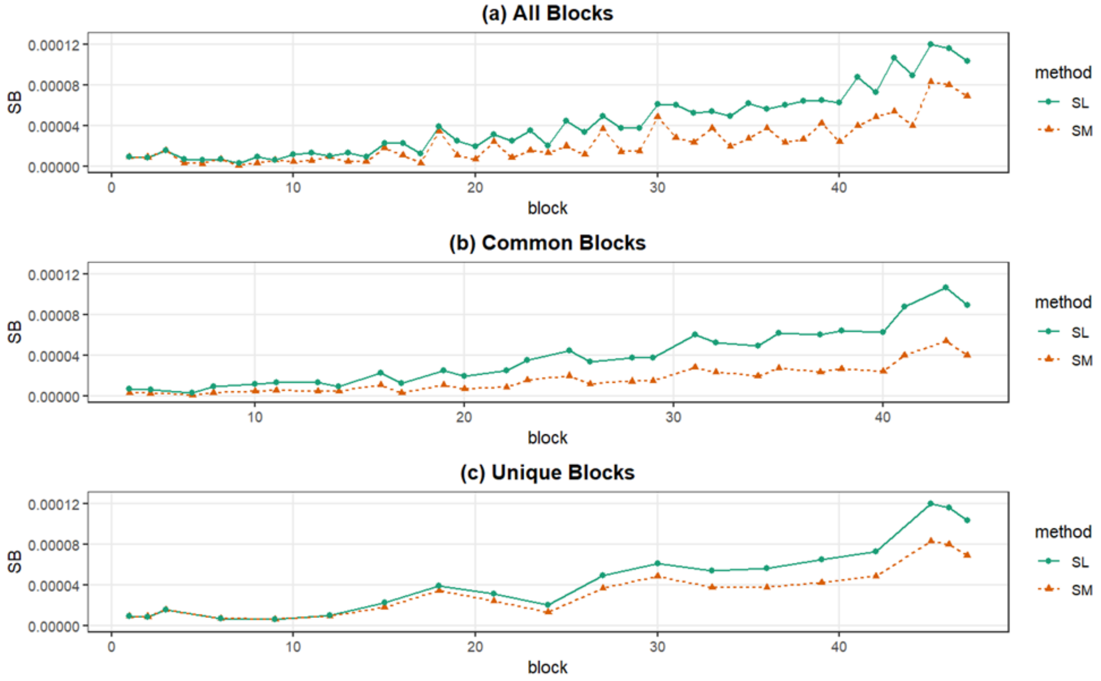


Figure 4

SB of Item Difficulty by Block (2-PL, Chained Design)

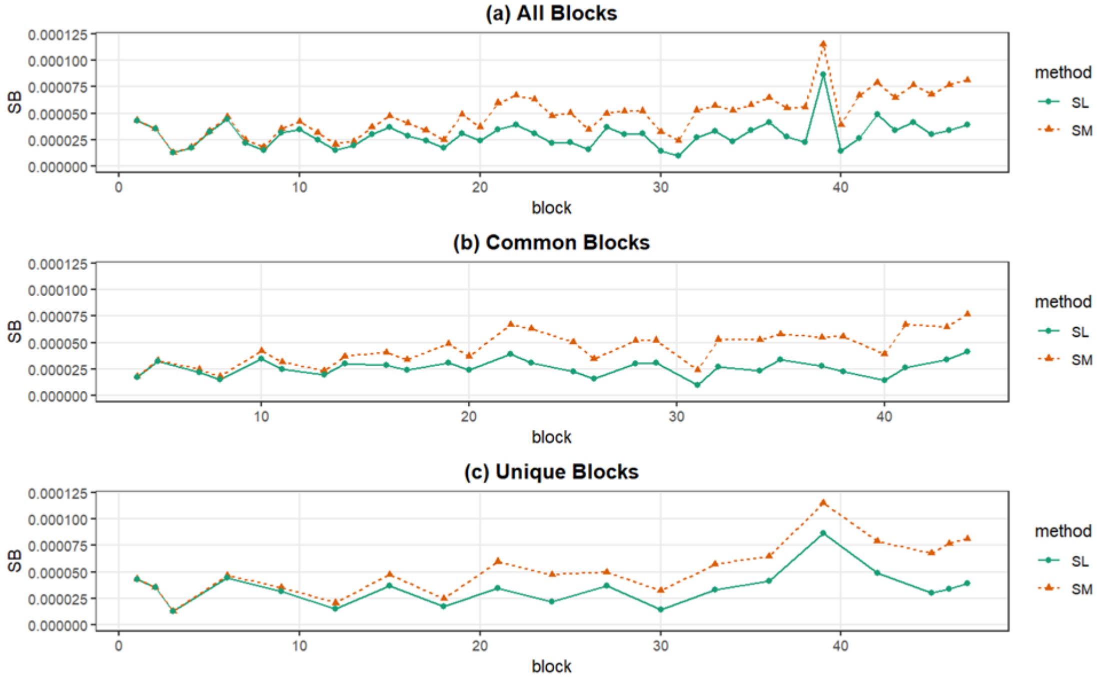


Figure 5
Variance of Item Discrimination by Block (2-PL, Chained Design)

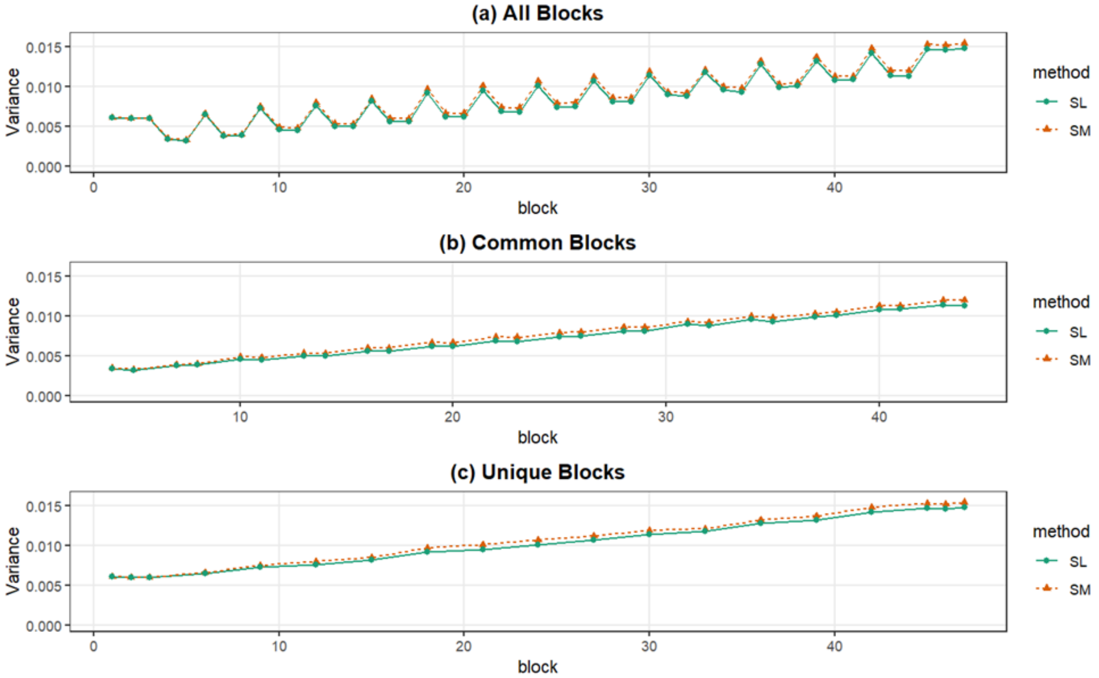


Figure 6
Variance of Item Difficulty by Block (2-PL, Chained Design)

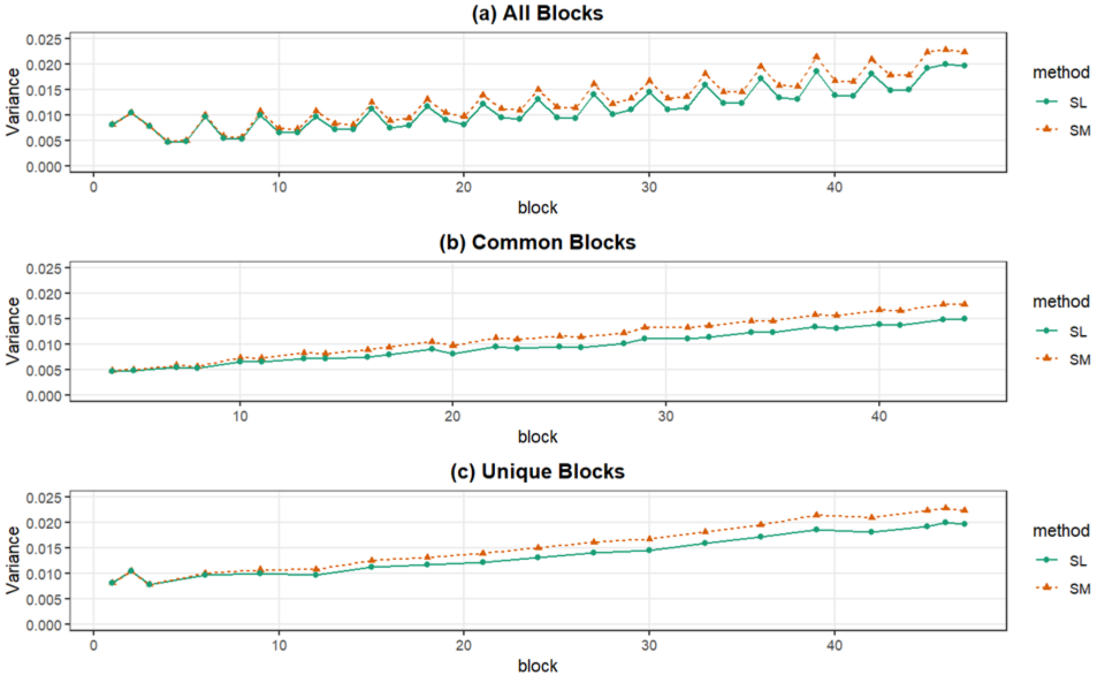


Figure 7

MSE of Item Discrimination by Block (2-PL, Chained Design)

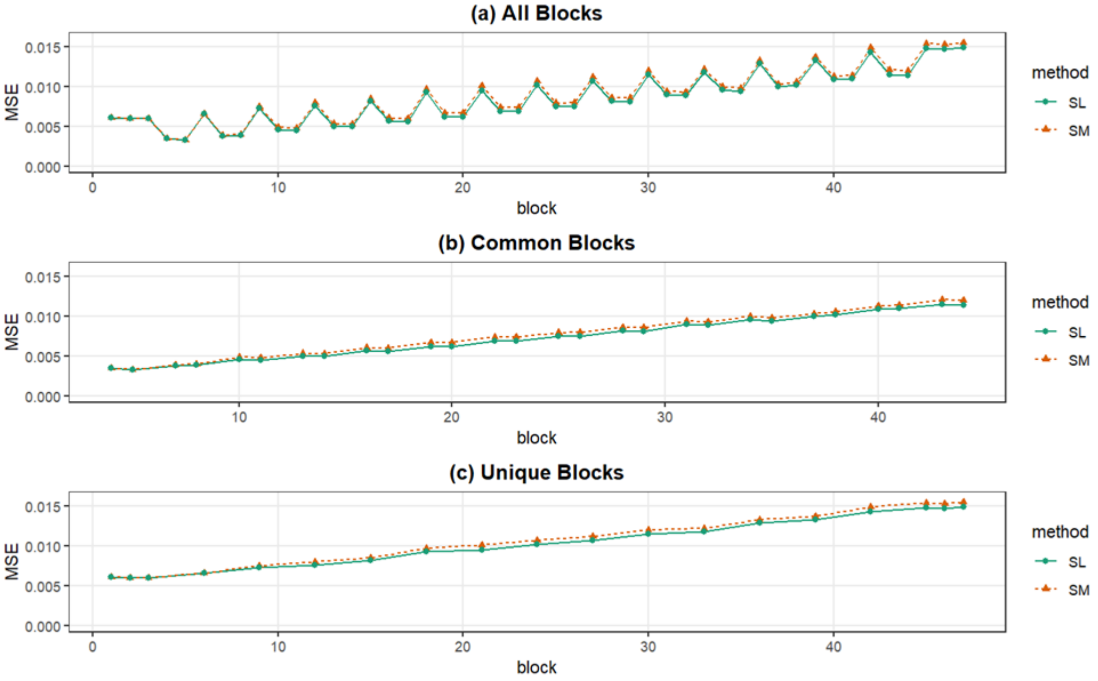


Figure 8

MSE of Item Difficulty by Block (2-PL, Chained Design)

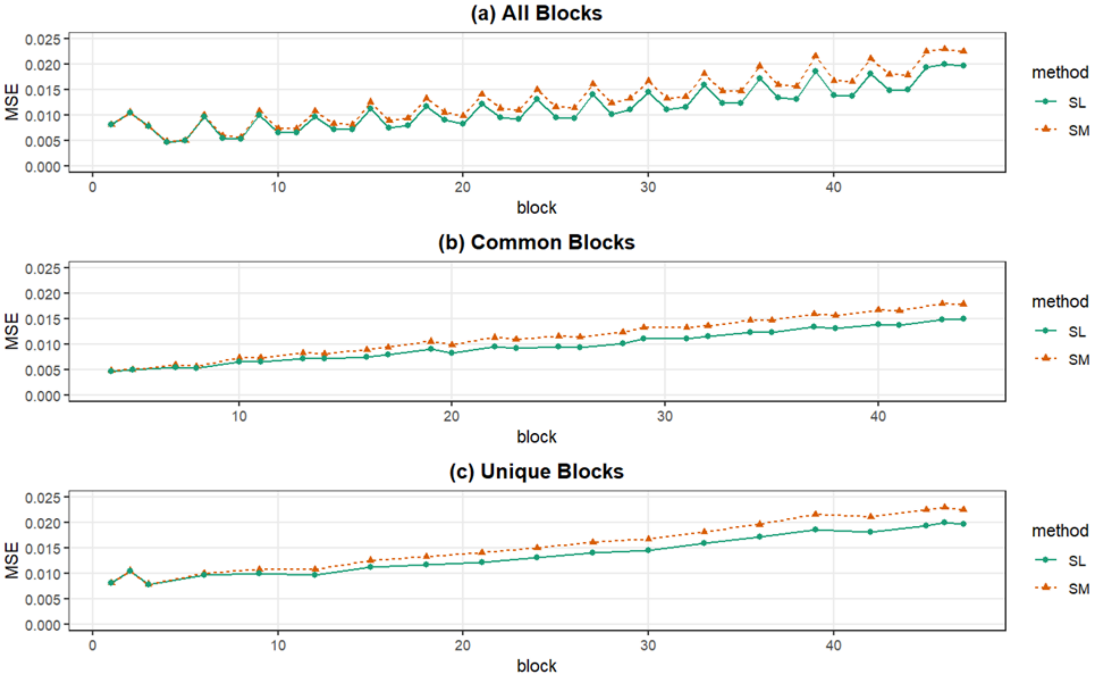


Figure 9

SB, Variance, MSE of Item Discrimination by Year (2-PL, Chained Design)

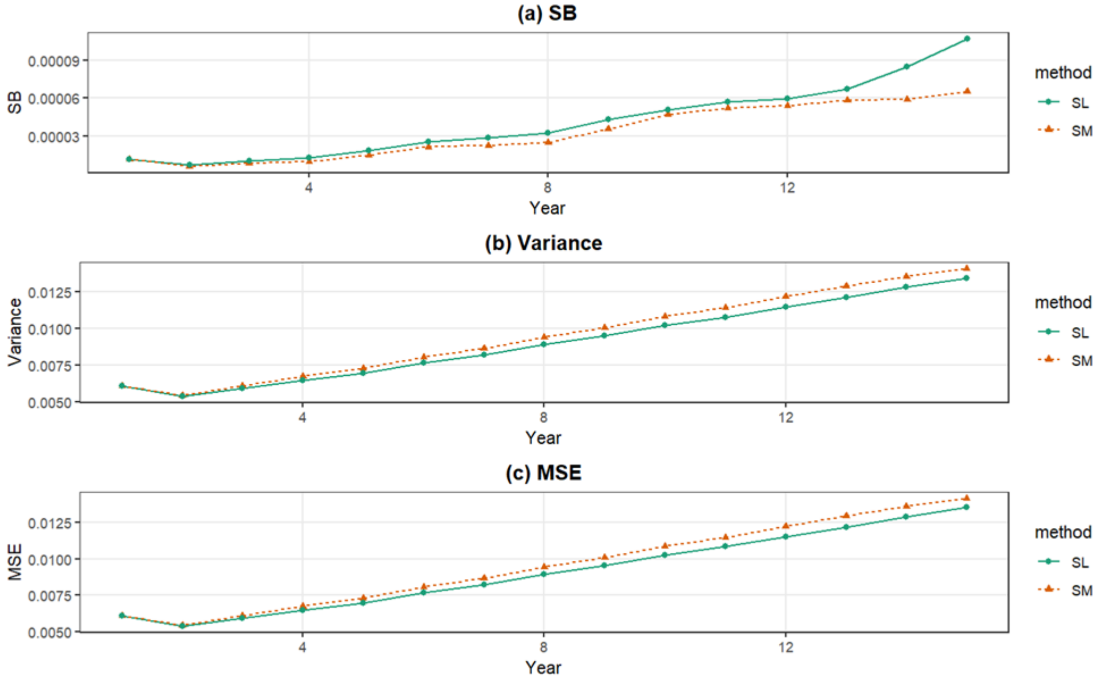


Figure 10

SB, Variance, MSE of Item Difficulty by Year (2-PL, Chained Design)

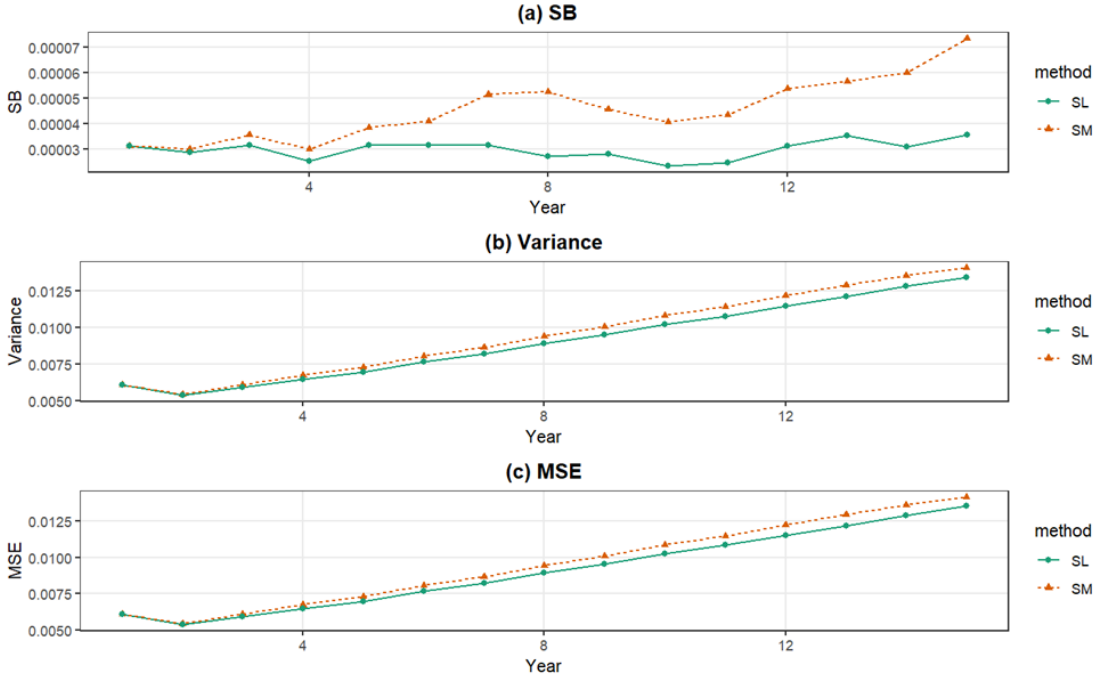


Figure 11

SB of TCC Criteria by Block (2-PL, Chained Design)

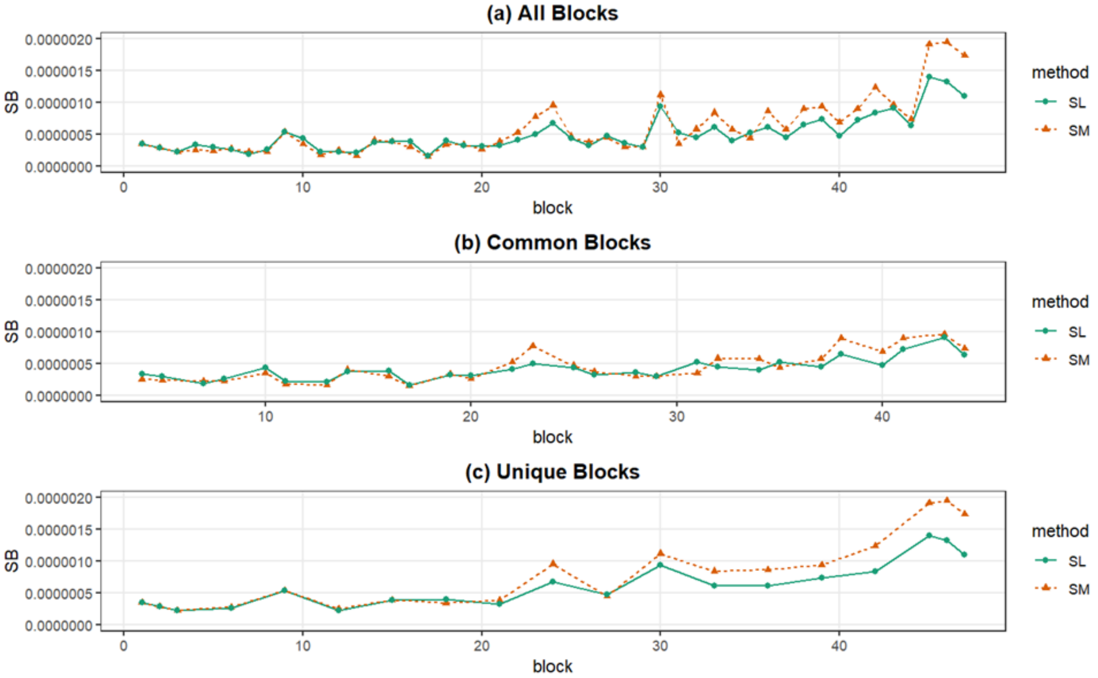


Figure 12

SB, Variance, MSE of TCC Criteria by Year (2-PL, Chained Design)

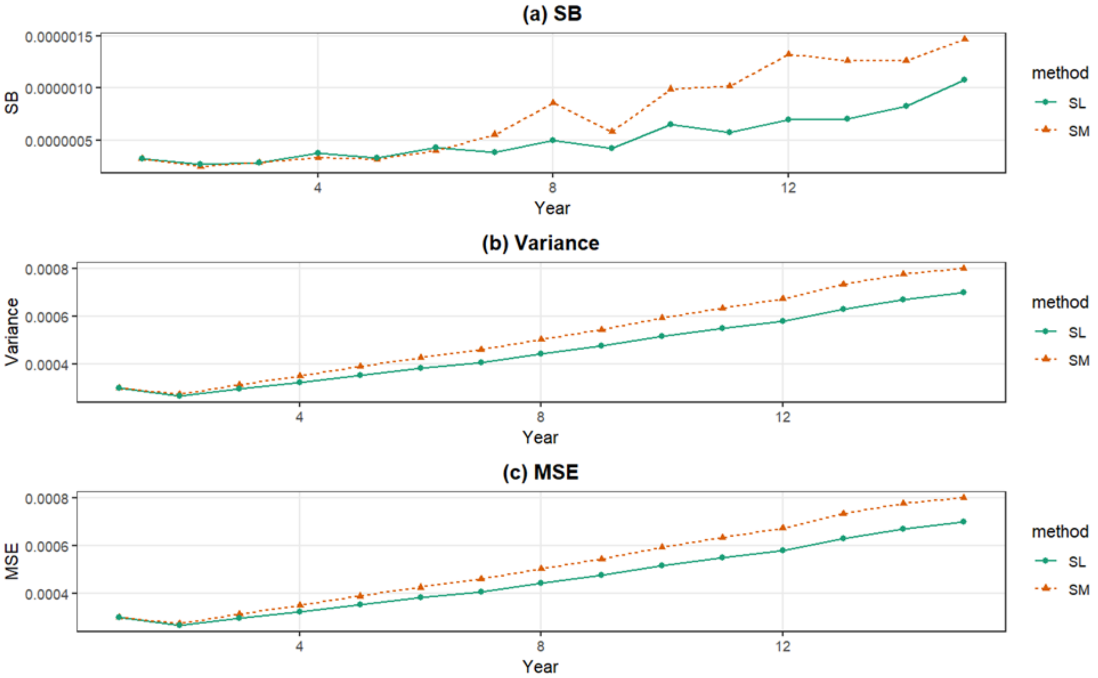


Figure 13

SB of Item Discrimination by Block (2-PL, 5-Replication Design)

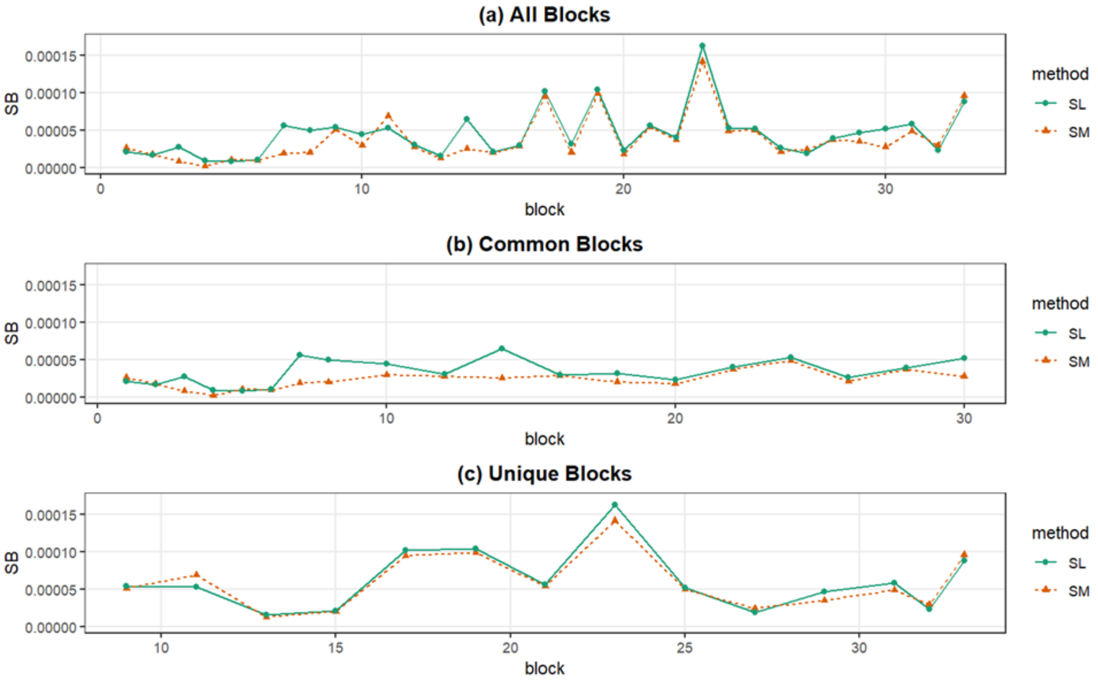


Figure 14

SB of Item Difficulty by Block (2-PL, 5-Replication Design)

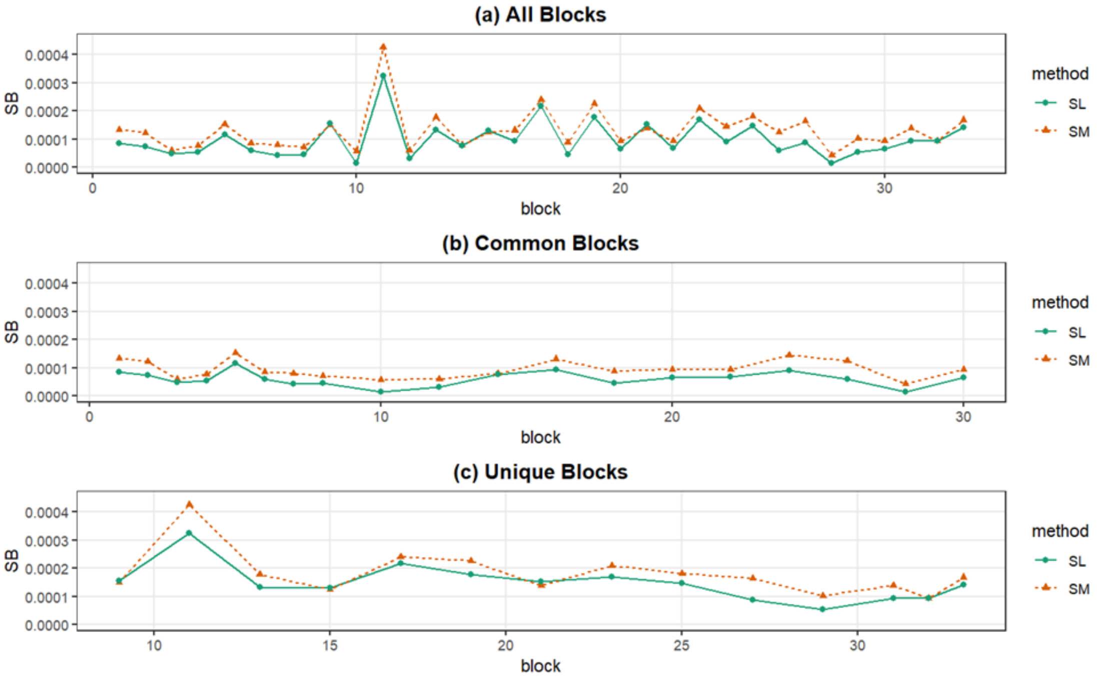


Figure 15

Variance of Item Discrimination by Block (2-PL, 5-Replication Design)

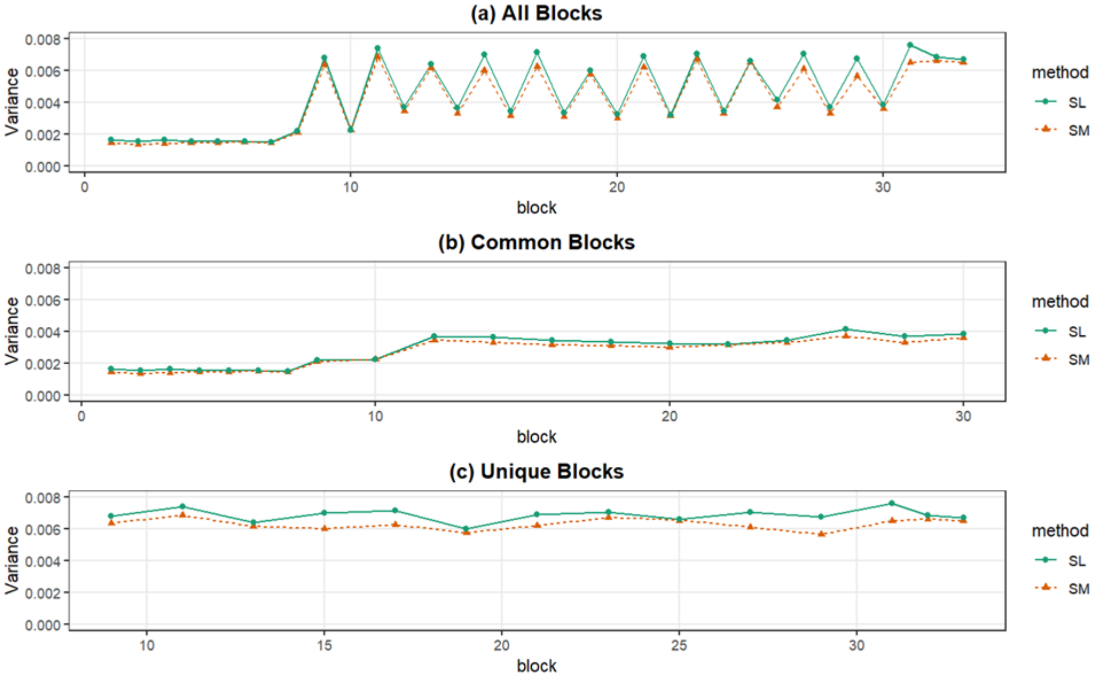


Figure 16

Variance of Item Difficulty by Block (2-PL, 5-Replication Design)

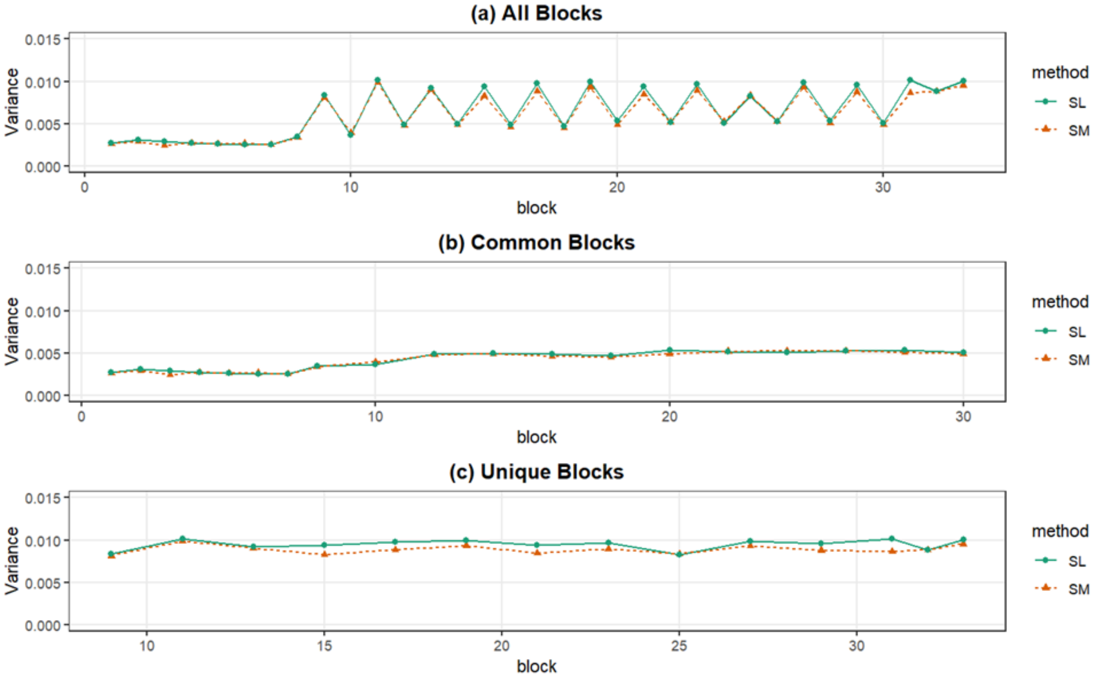


Figure 17

MSE of Item Discrimination by Block (2-PL, 5-Replication Design)

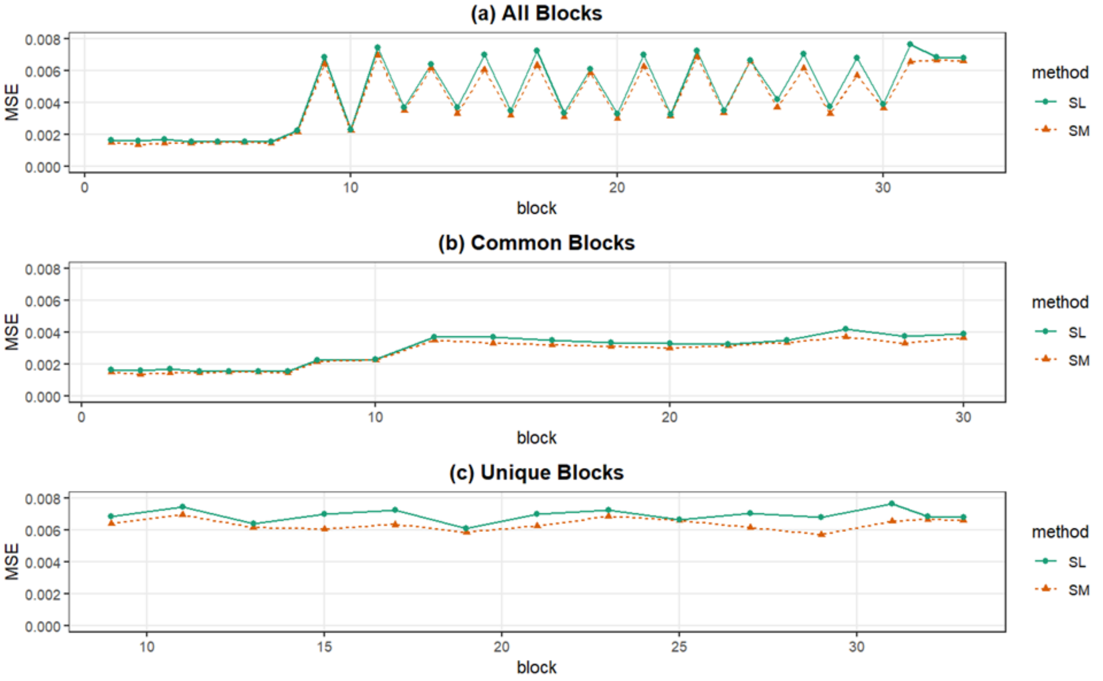


Figure 18

MSE of Item Difficulty by Block (2-PL, 5-Replication Design)

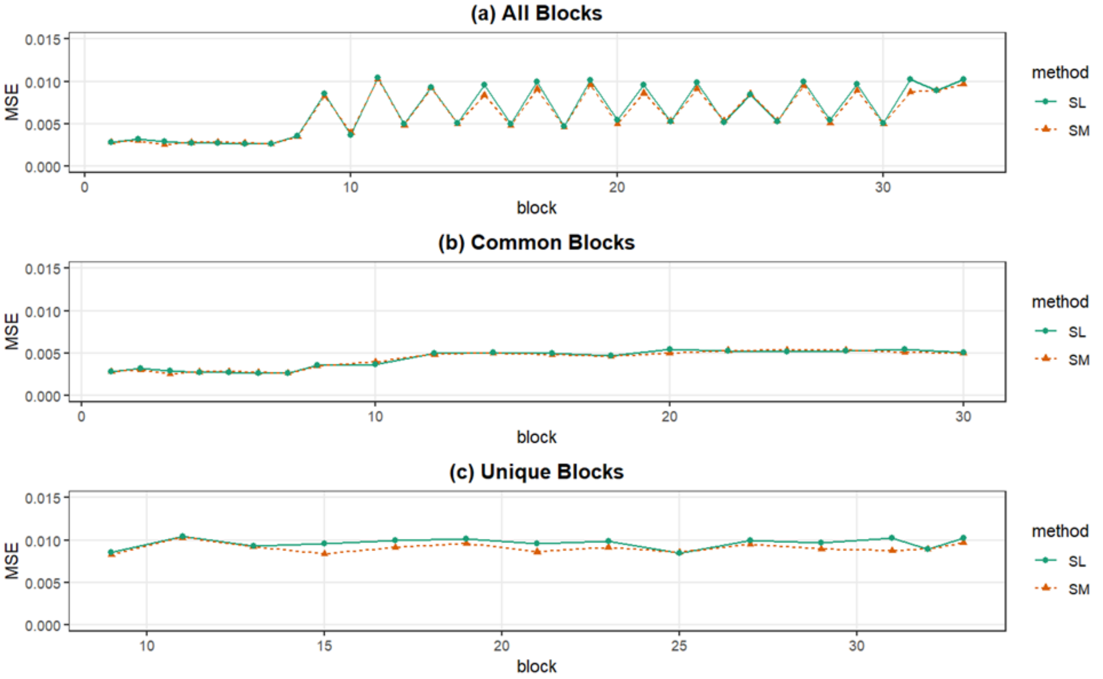


Figure 19

SB, Variance, MSE of Item Discrimination by Year (2-PL, 5-Replication Design)

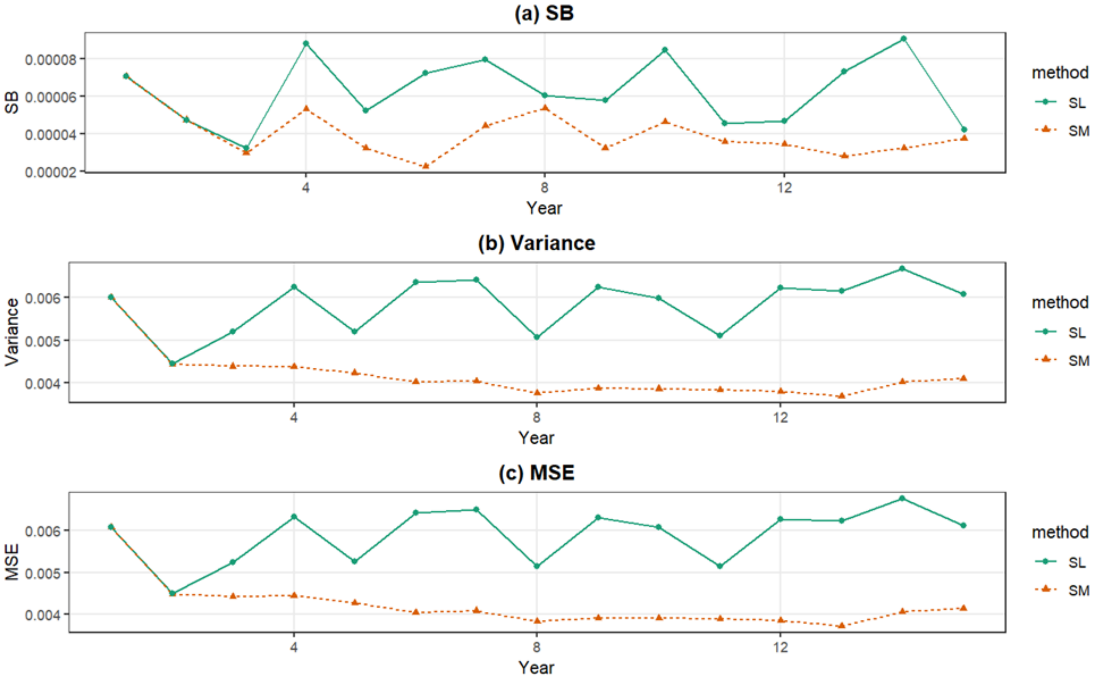


Figure 20

SB, Variance, MSE of Item Difficulty by Year (2-PL, 5-Replication Design)

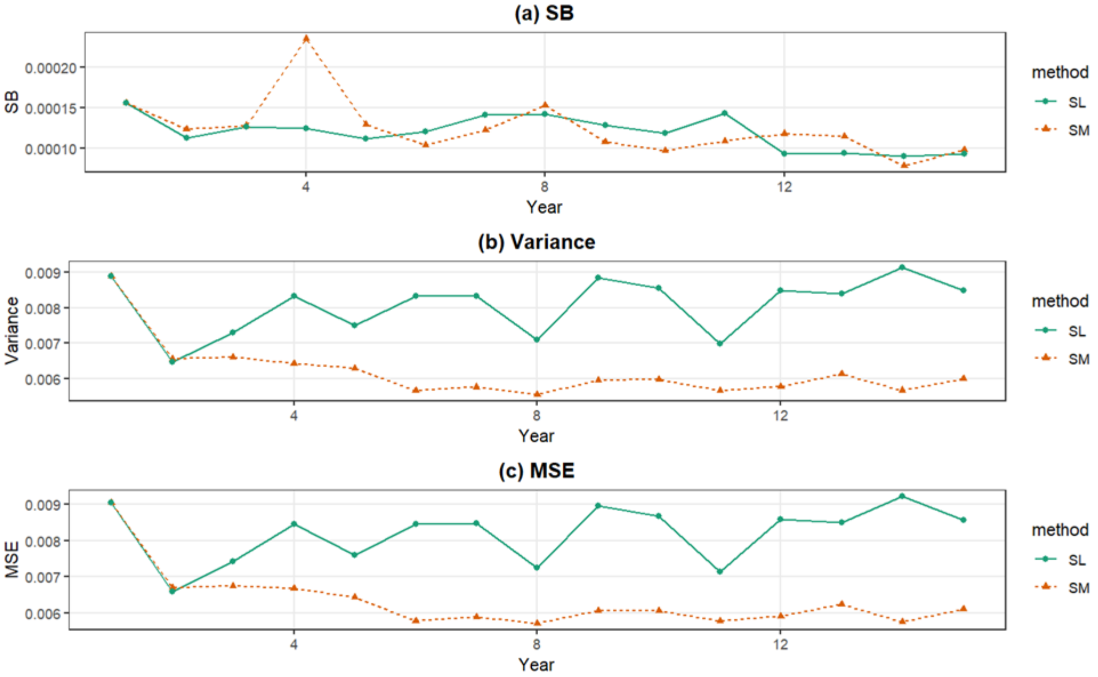


Figure 21

SB of TCC Criteria by Block (2-PL, 5-Replication Design)

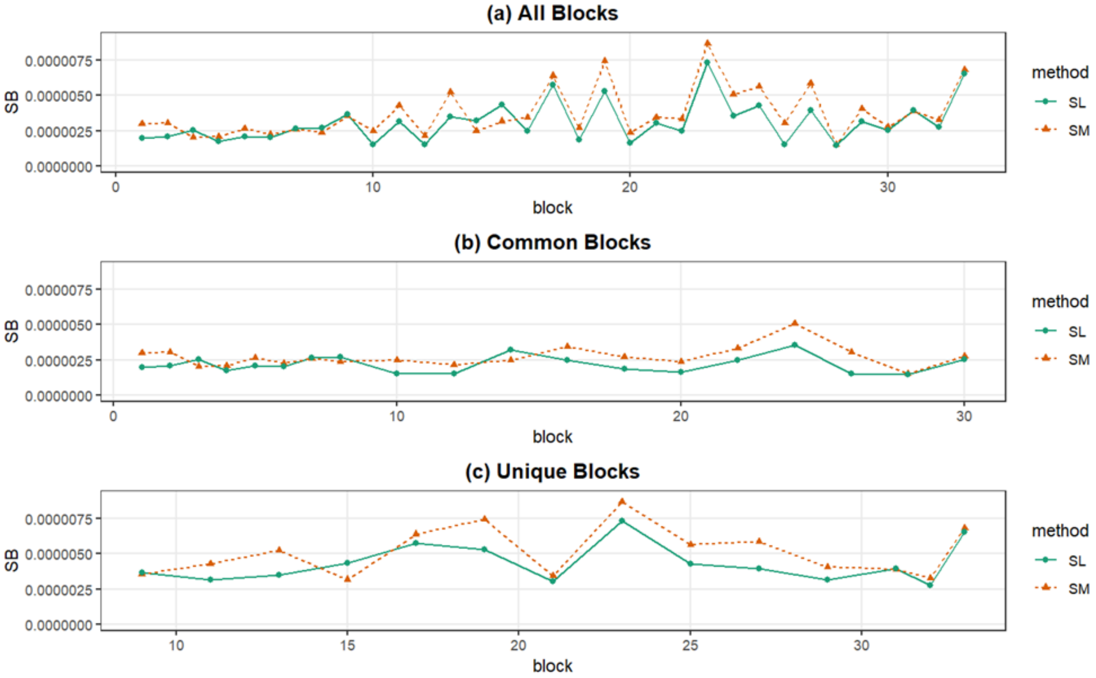


Figure 22

SB, Variance, MSE of TCC Criteria by Year (2-PL, 5-Replication Design)

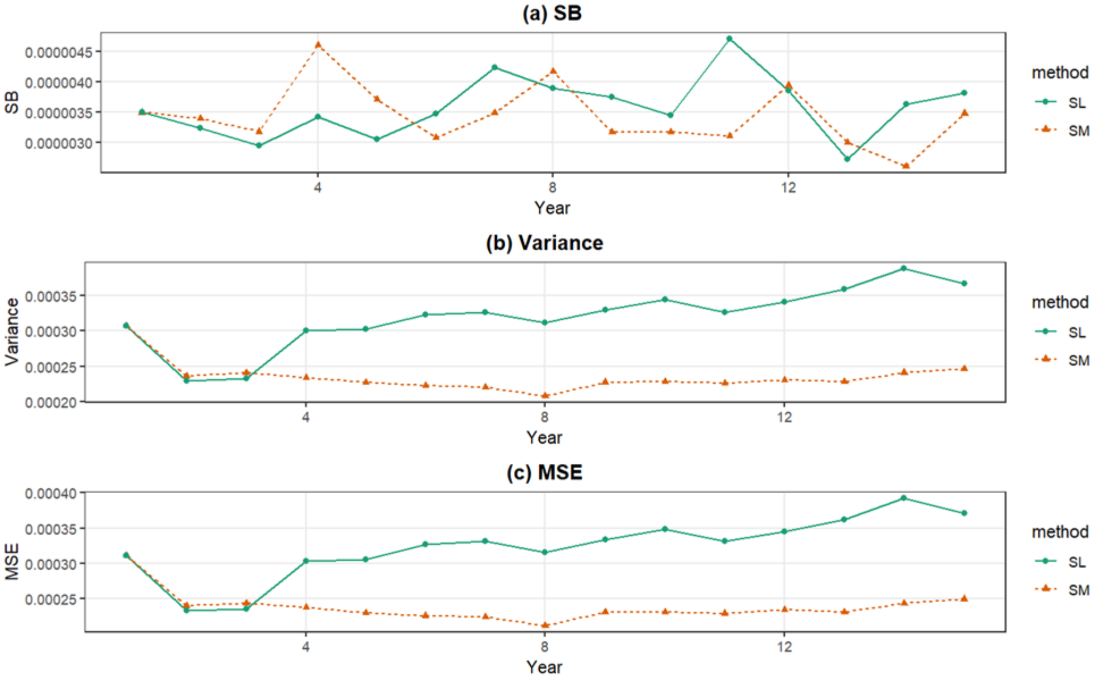


Figure 23

SB of Item Discrimination for the First Seven Blocks (2-PL, 5-Replication Design)

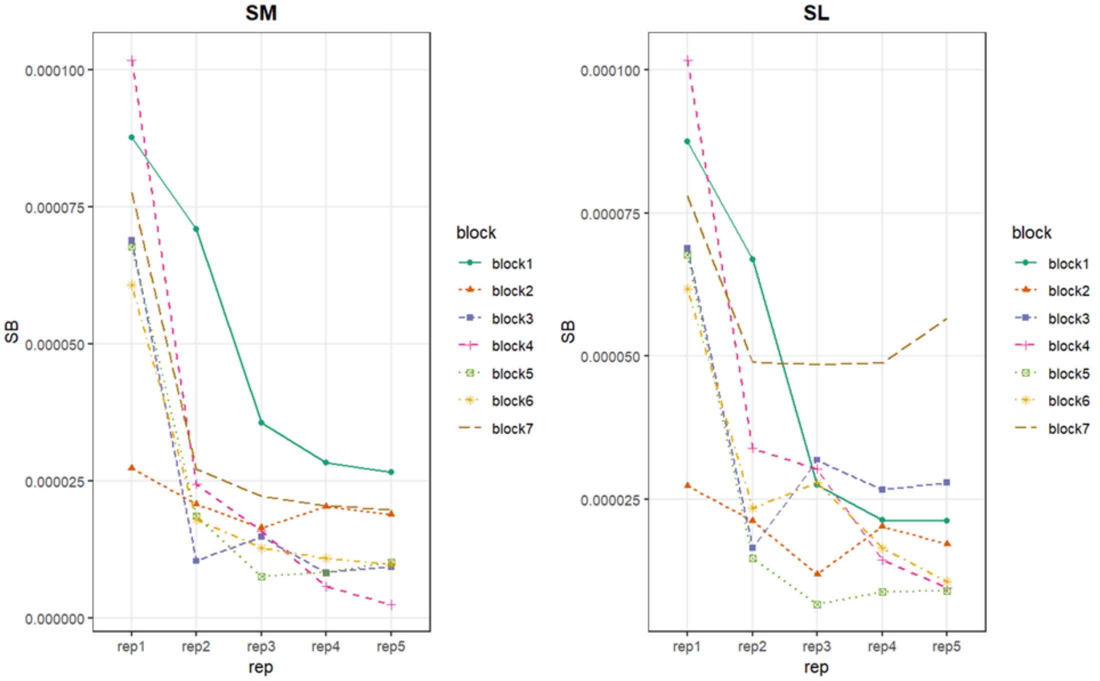


Figure 24

SB of Item Difficulty for the First Seven Blocks (2-PL, 5-Replication Design)

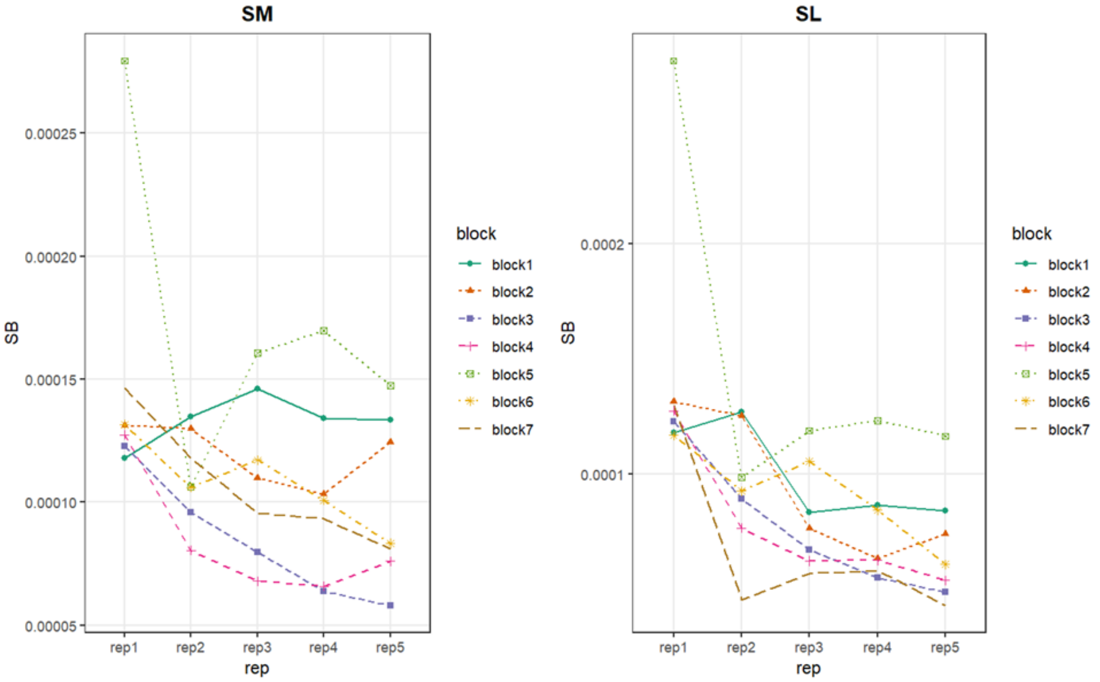


Figure 25

SB of Item Difficulty by Block (1-PL, Chained Design)

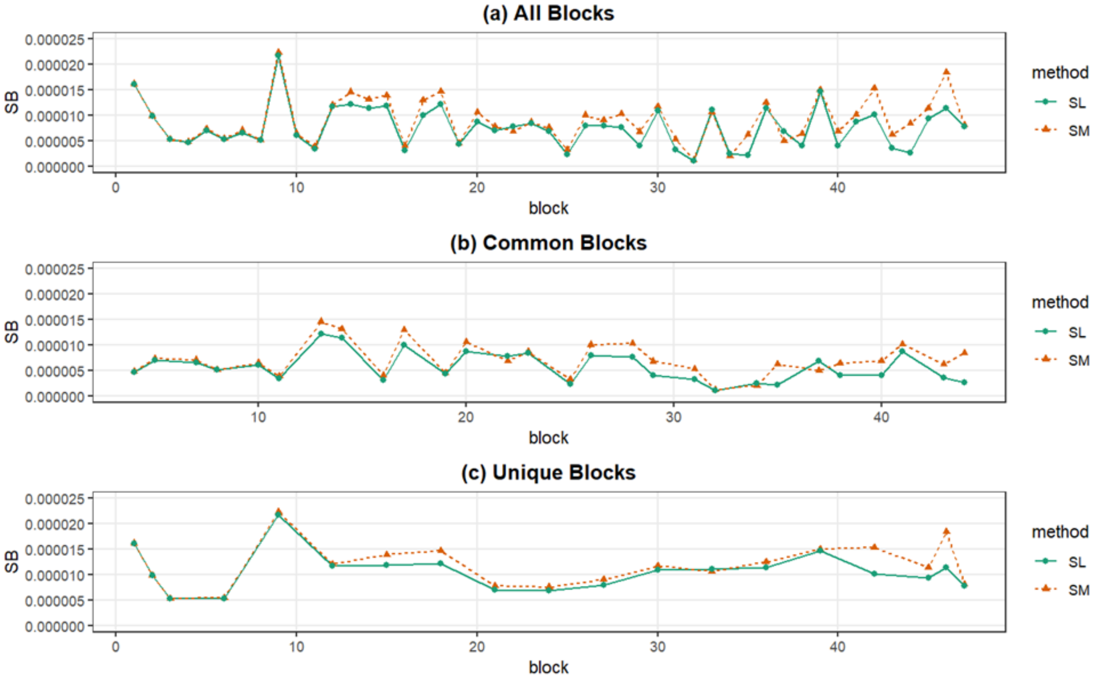


Figure 26

Variance of Item Difficulty by Block (1-PL, Chained Design)

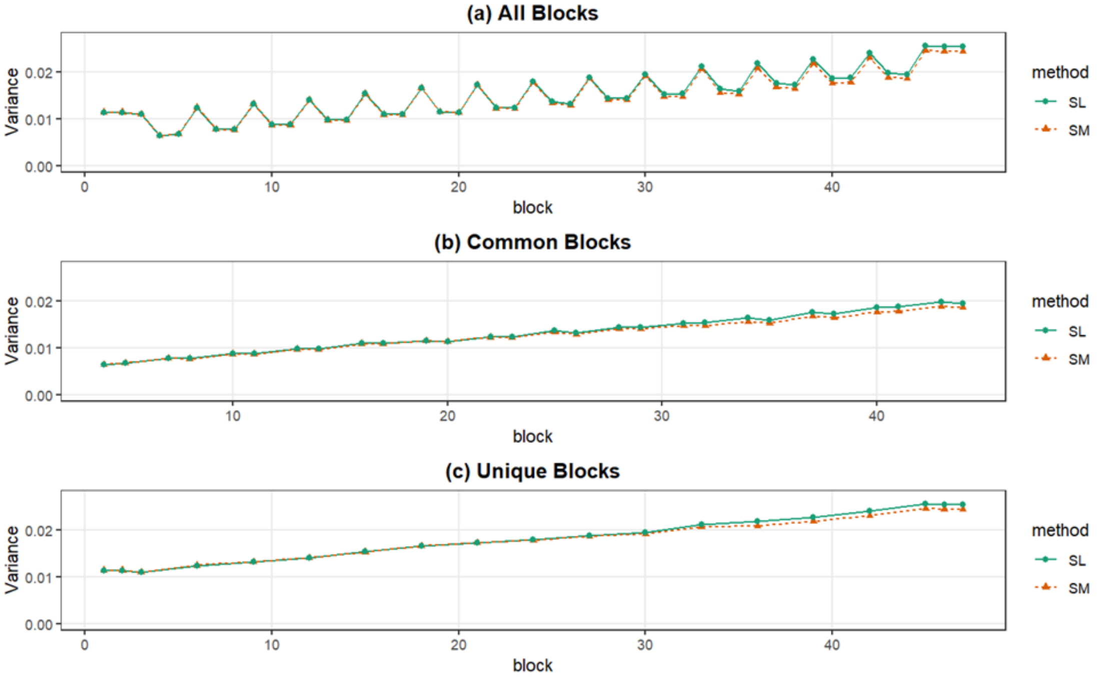


Figure 27
MSE of Item Difficulty by Block (1-PL, Chained Design)

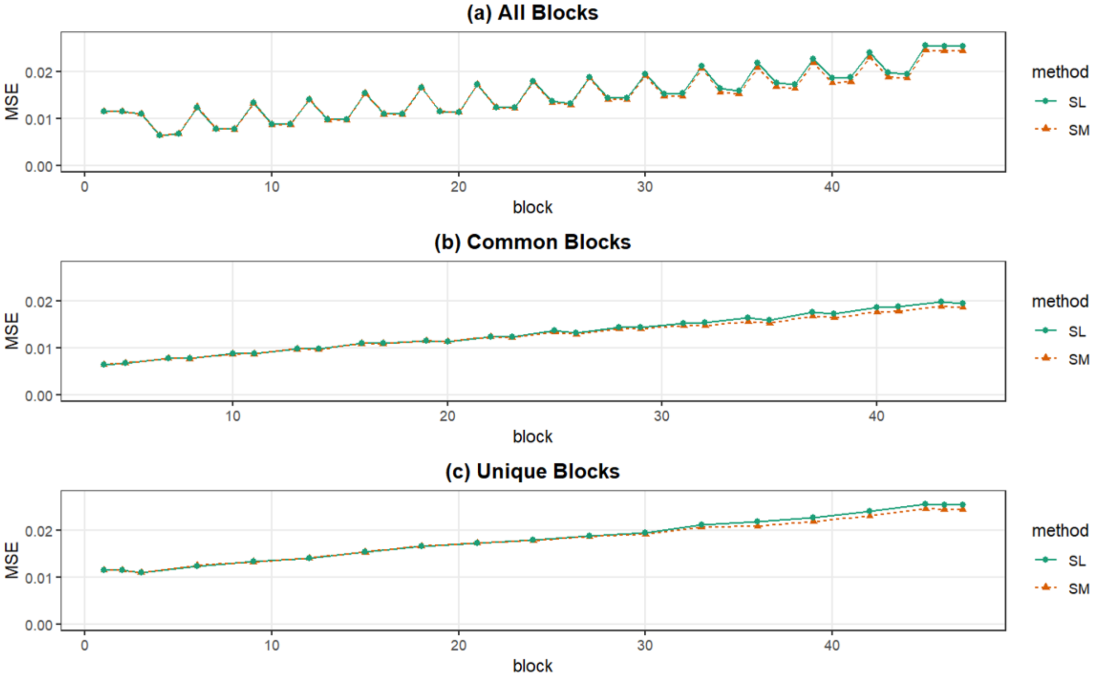


Figure 28
SB, Variance, MSE of Item Difficulty by Year (1-PL, Chained Design)

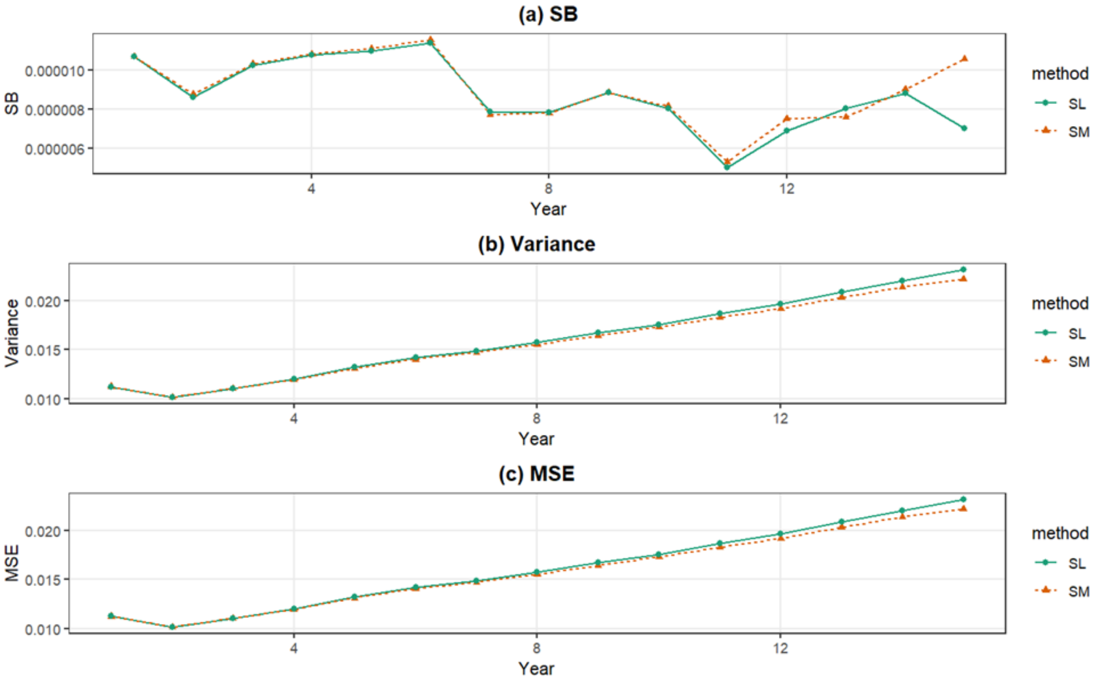


Figure 29

SB of TCC Criteria by Block (1-PL, Chained Design)

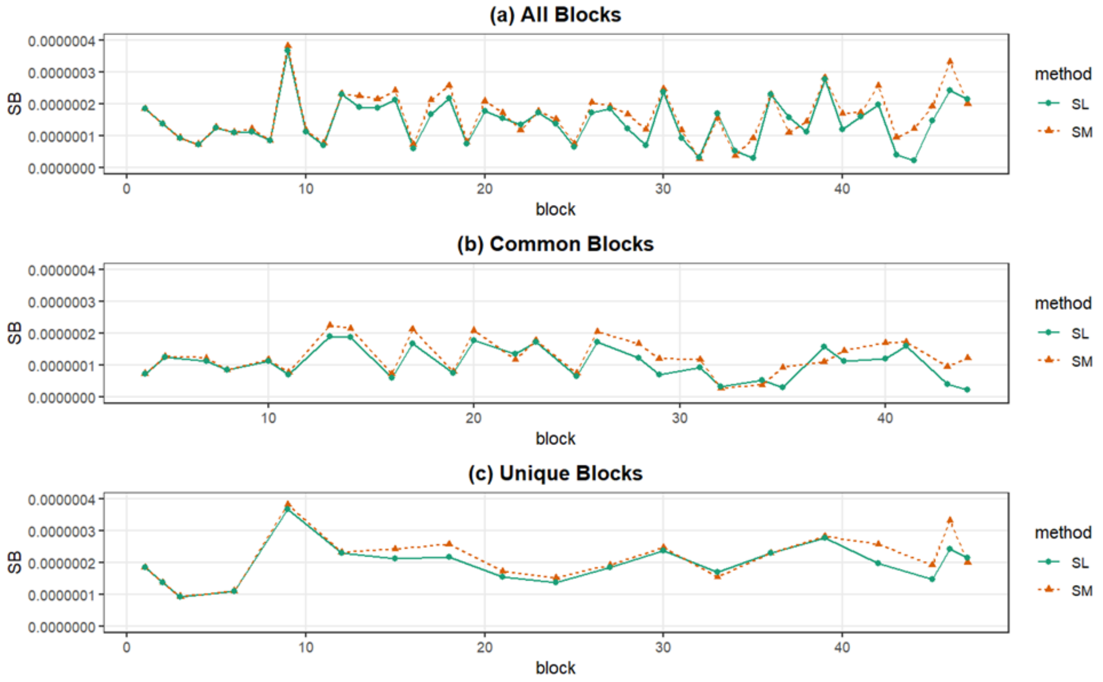


Figure 30

SB, Variance, MSE of TCC Criteria by Year (1-PL, Chained Design)

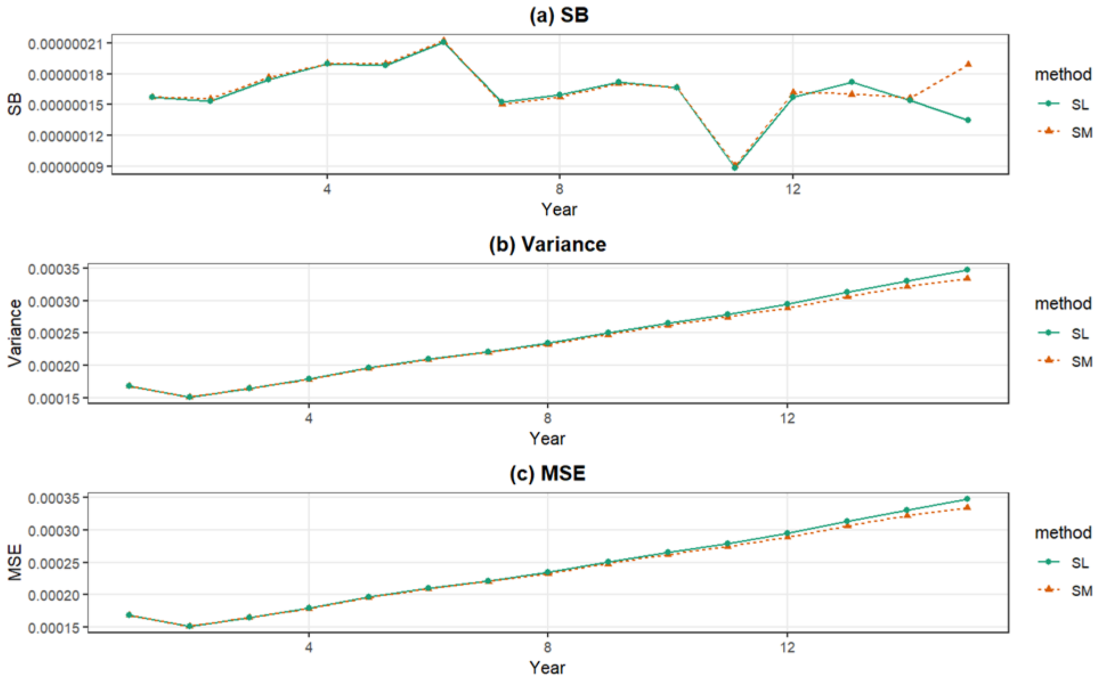


Figure 31

SB of Item Difficulty by Block (1-PL, 5-Replication Design)

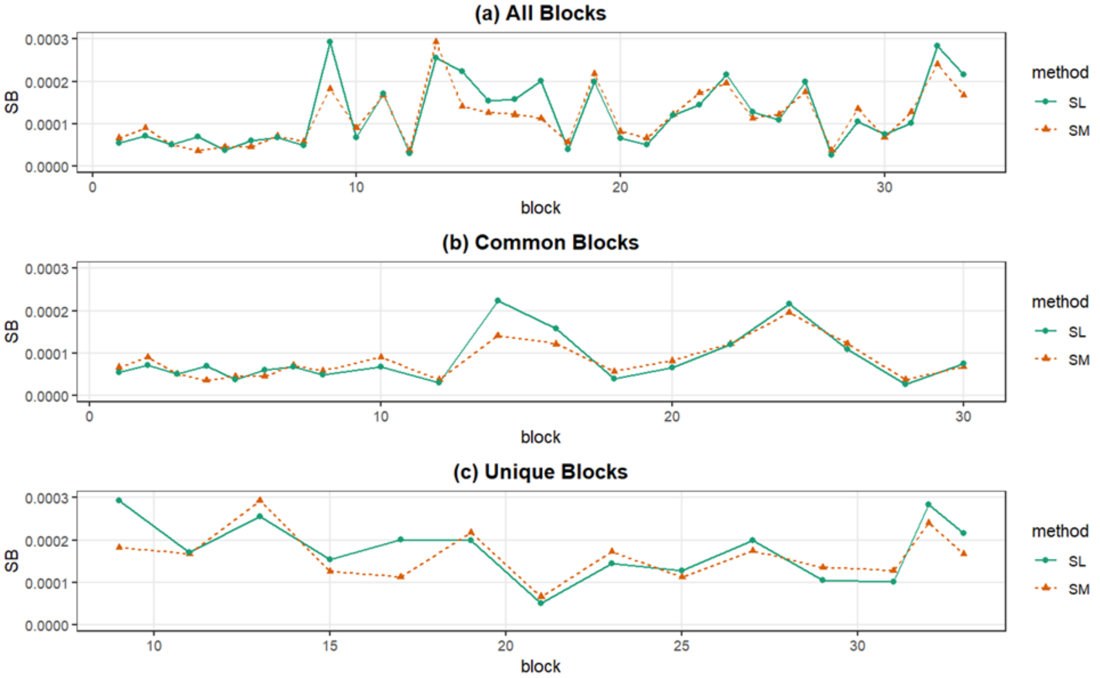


Figure 32

Variance of Item Difficulty by Block (1-PL, 5-Replication Design)

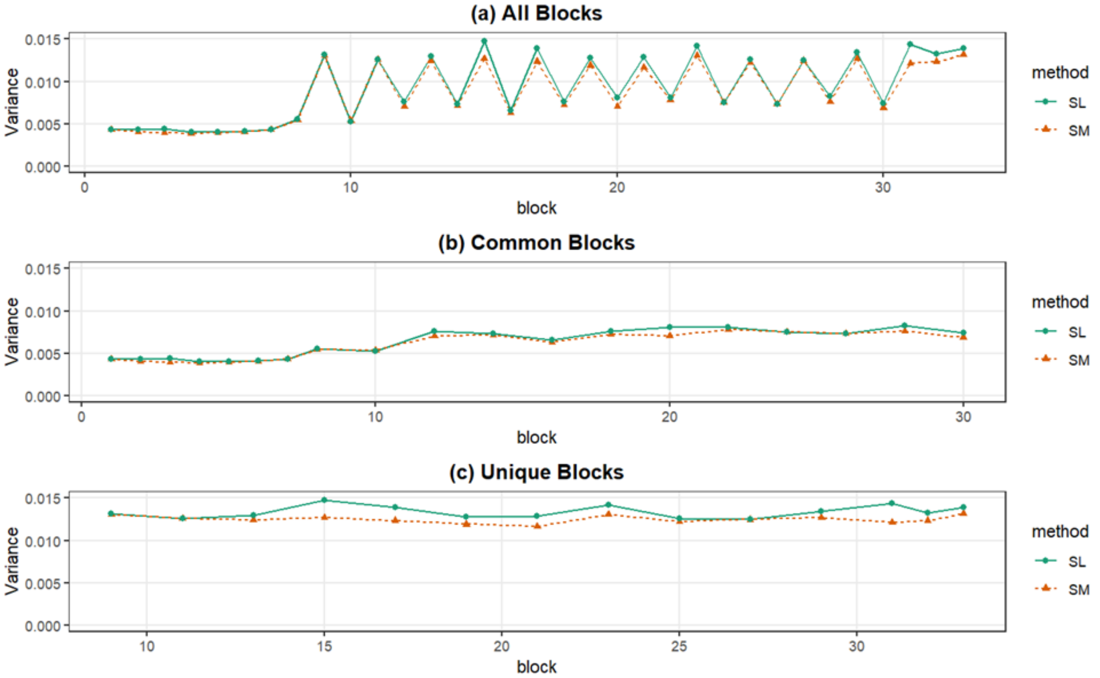


Figure 33

MSE of Item Difficulty by Block (1-PL, 5-Replication Design)

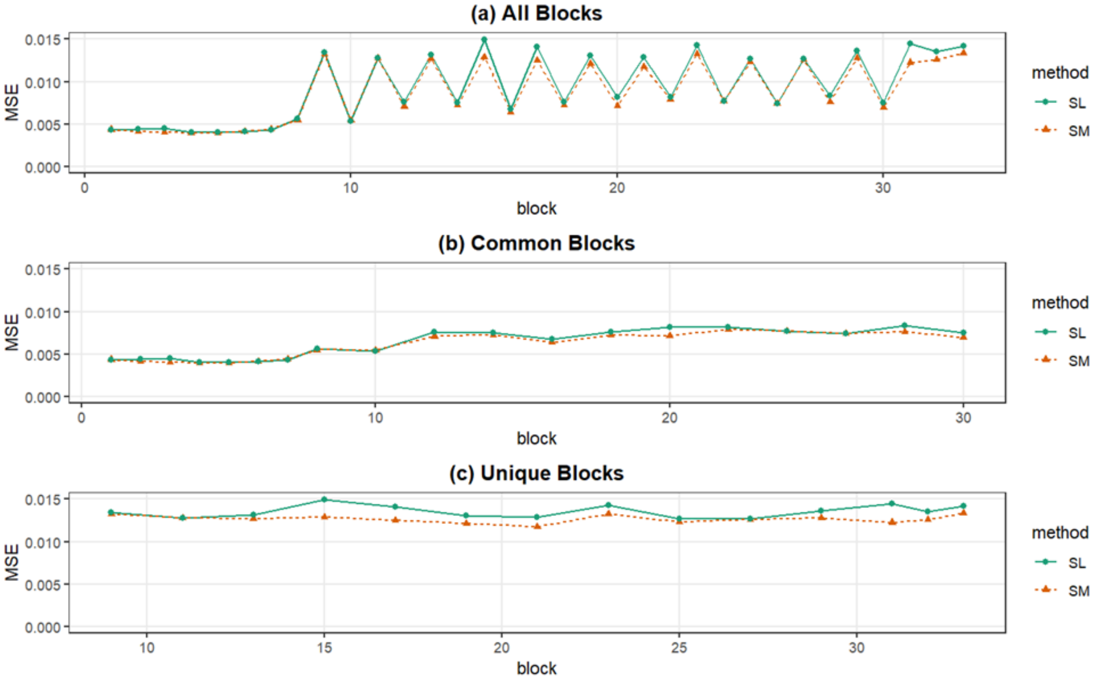


Figure 34

SB, Variance, MSE of Item Difficulty by Year (1-PL, 5-Replication Design)

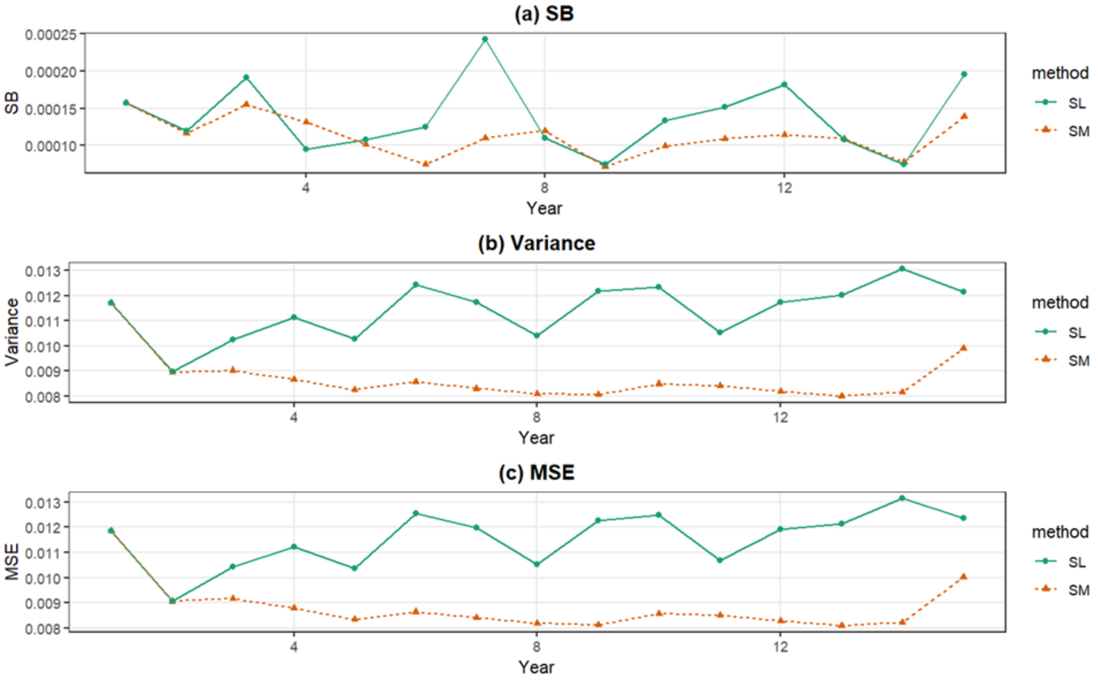


Figure 35

SB of TCC Criteria by Block (1-PL, 5-Replication Design)

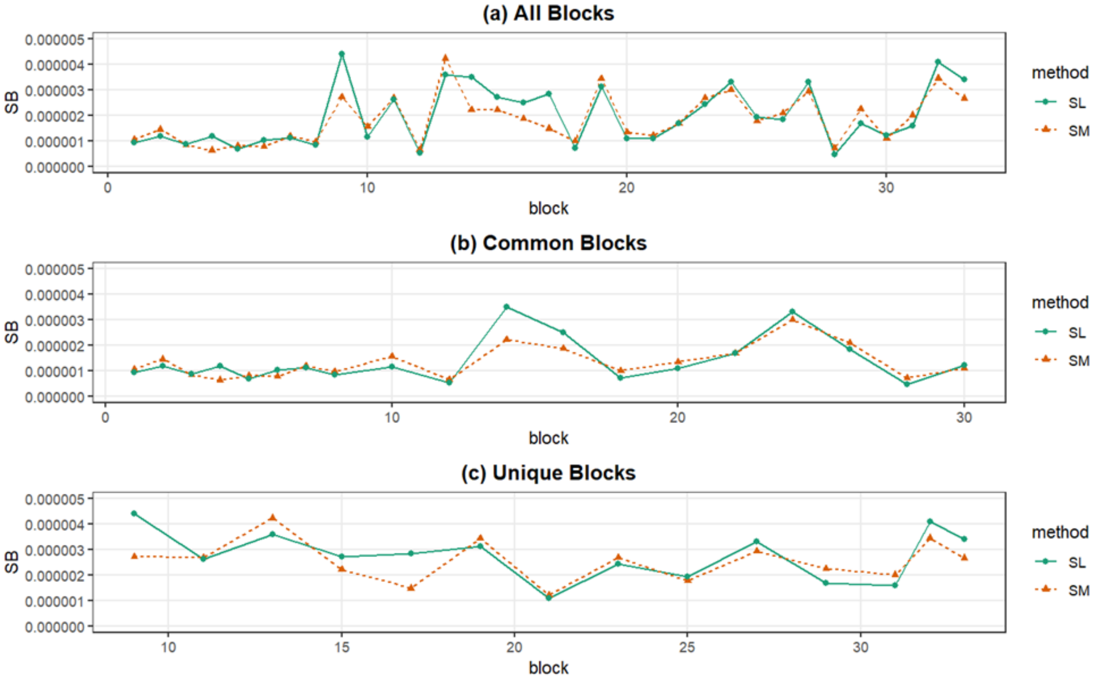


Figure 36

SB, Variance, MSE of TCC Criteria by Year (1-PL, 5-Replication Design)

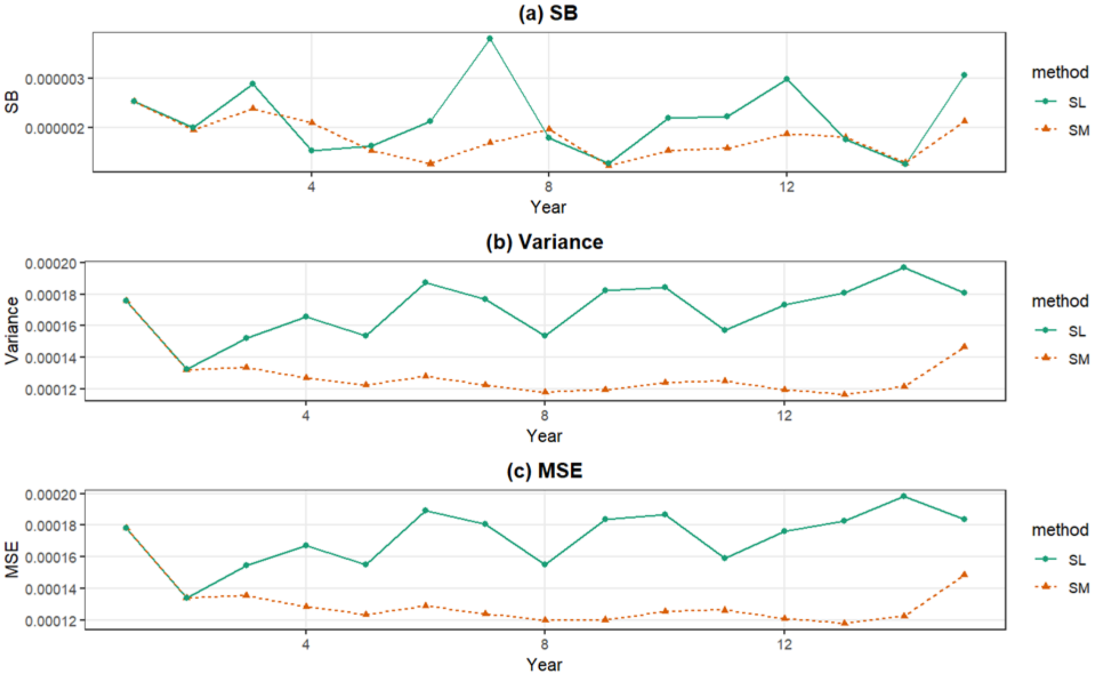


Figure 37

SB of Item Difficulty for the First Seven Blocks (1-PL, 5-Replication Design)

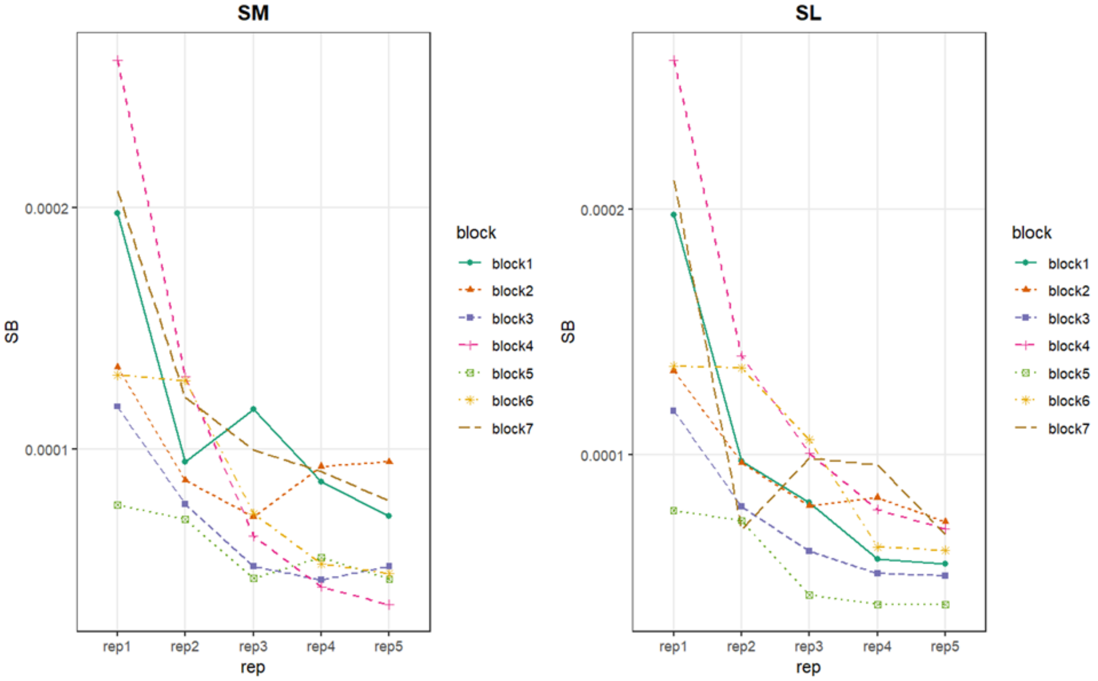


Figure 38

Relationship between True Item Discrimination and SB of the Estimated Item Discrimination (2-PL, Chained Design)

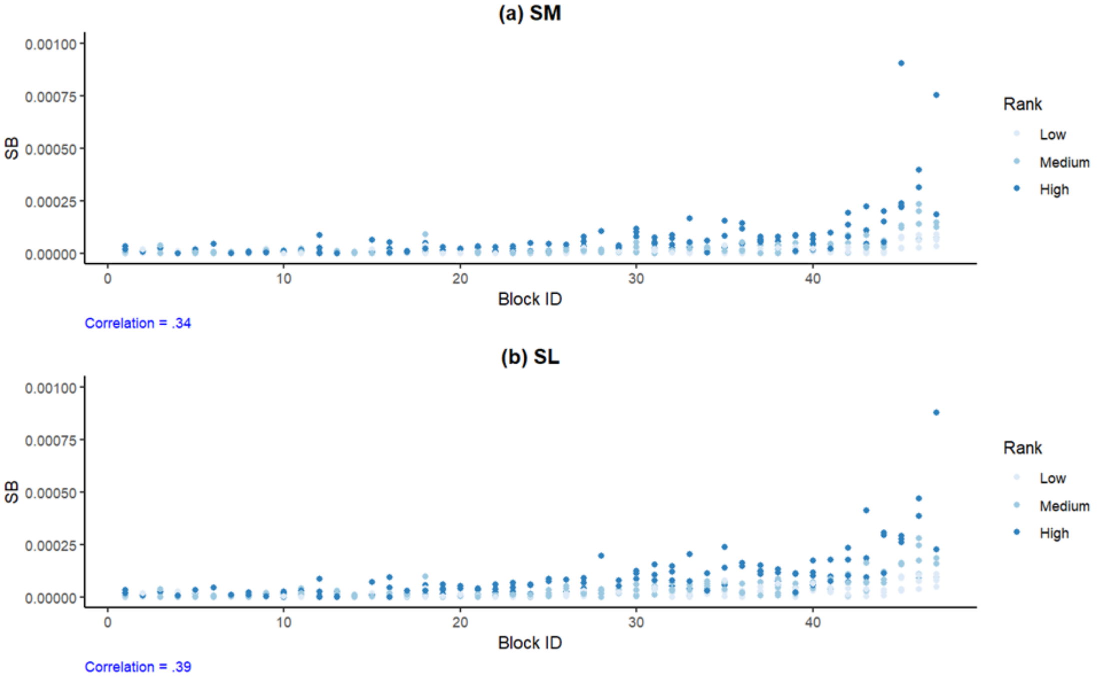


Figure 39
Relationship between True Item Difficulty and SB of the Estimated Item Difficulty (2-PL, Chained Design)

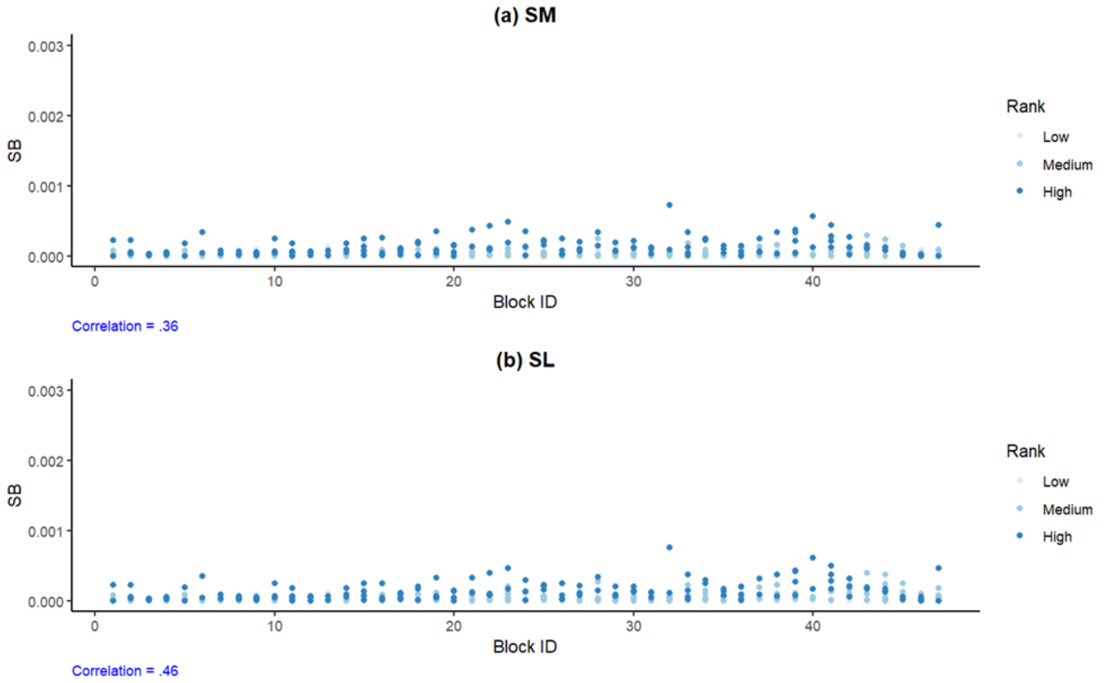


Figure 40
Relationship between True Item Discrimination and Variance of the Estimated Item Discrimination (2-PL, Chained Design)

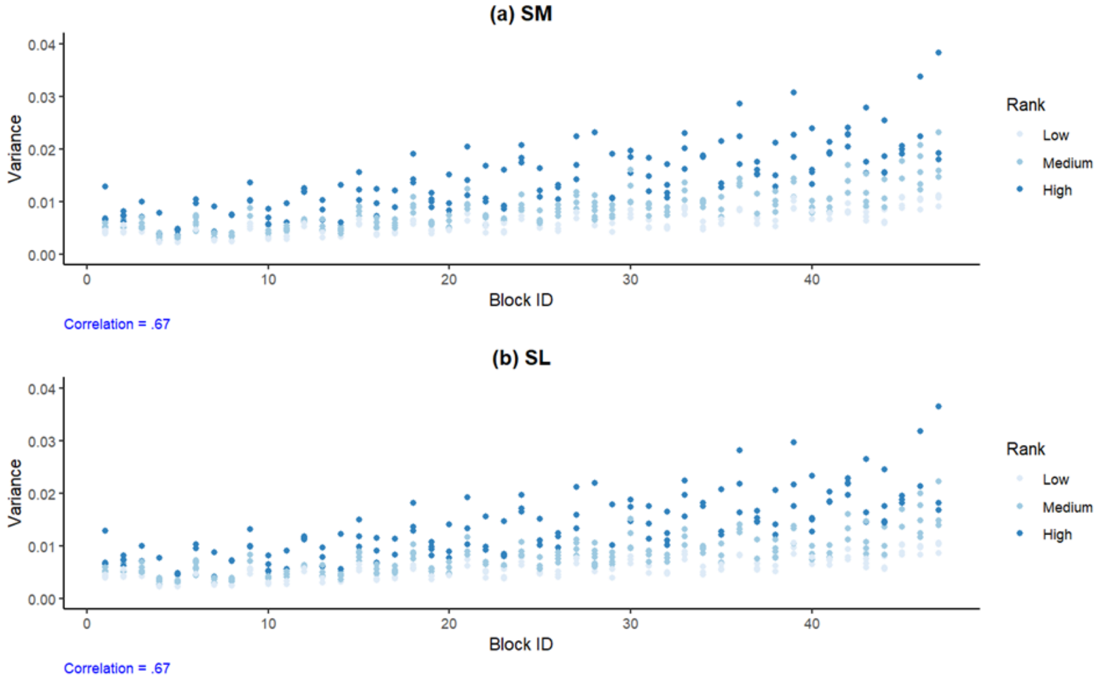


Figure 41
Relationship between True Item Difficulty and Variance of the Estimated Item Difficulty (2-PL, Chained Design)

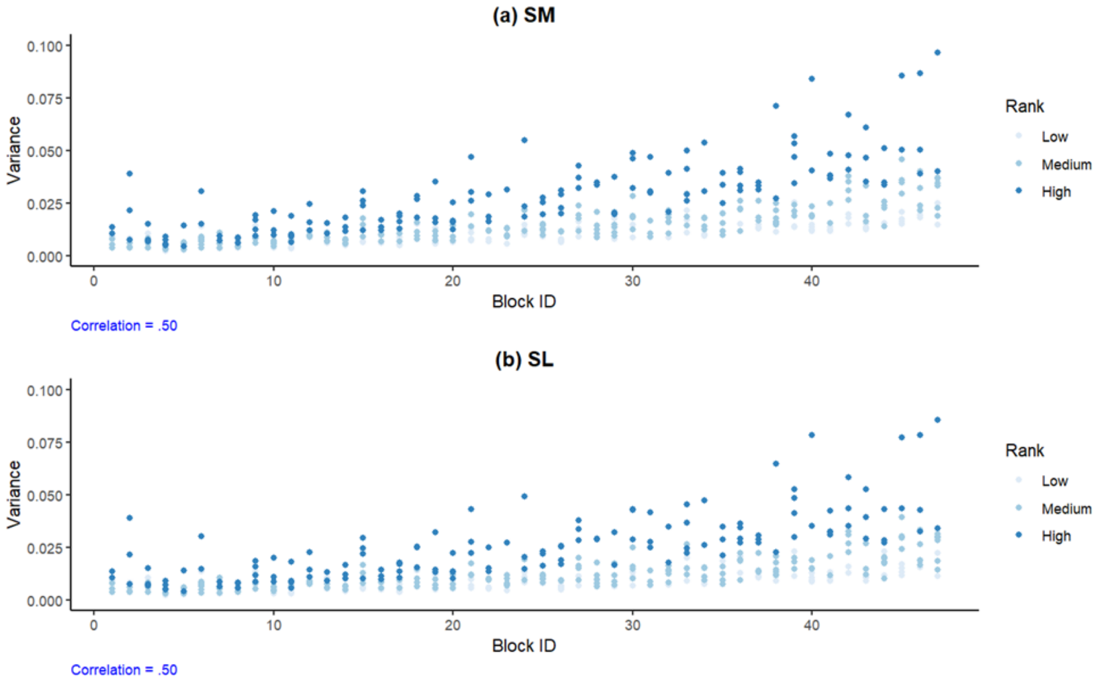


Figure 42
Relationship between True Item Discrimination and MSE of the Estimated Item Discrimination (2-PL, Chained Design)

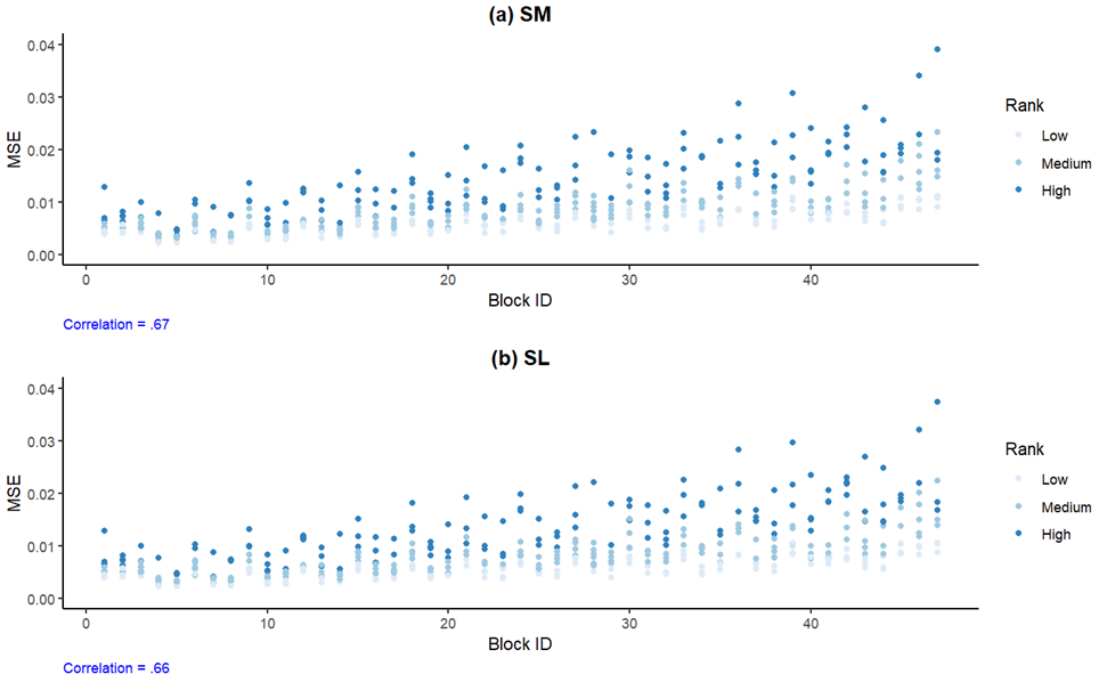
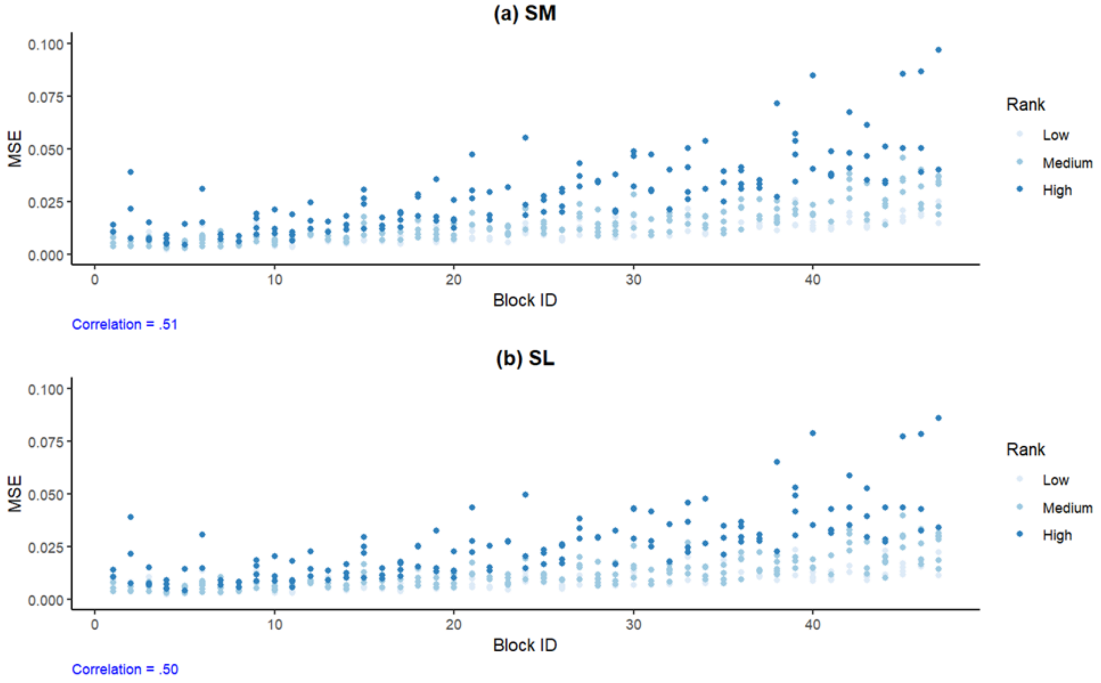


Figure 43

Relationship between True Item Difficulty and MSE of the Estimated Item Difficulty (2-PL, Chained Design)



(a) SM

(b) SL

References

Battauz, M. (2017). Multiple equating of separate IRT calibrations. *Psychometrika*, *82*(3), 610–636.

Battauz, M., & Leôncio, W. (2023). A likelihood approach to item response theory equating of multiple forms. *Applied Psychological Measurement*, *47*(3), 200–220.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, *28*(4), 3–14.

Béguin, A. A., & Hanson, B. A. (2001, March). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating.* Papaer presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Haberman, S. J. (2009). *Linking parameter estimates derived from an item response model through separate calibrations* (ETS Research Report Series No. 09-40*).* Princeton, NJ: Educational Testing Service.

Haberman, S. J. (2010). *Limits on the accuracy of linking* (ETS Research Report Series No. 10-22*).* Princeton, NJ: Educational Testing Service.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*(3), 144–149.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(1), 3–24.

Kang, T., & Petersen, N. S. (2012). Linking item parameters to a base scale. *Asia Pacific Education Review*, *13*(2), 311–321.

Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied psychological measurement*, *22*(2), 131–143.

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

Lee, W., & Ban, J. C. (2009). A comparison of IRT linking procedures. *Applied Measurement in Education*, *23*(1), 23–48.

Lee, W., & Lee, G. (2018). IRT equating and linking. In P. Irwing, T. Booth, D. J. Hughes (Eds.)*, The wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development.* London: John Wiley & Sons.

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

Lu, Y., & Antal, J. (2022, April). *Simultaneous linking for improving scale stability.* Paper presented at the Annual meeting of the National Council on Measurement in Education, San Diego, CA.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Sientific Software.

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in psychology*, *5*(978).

Partchev, I., Maris, G., & Hattori, T. (2022). irtoys: A collection of functions related to item response theory (IRT) [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=irtoys` (R package version 0.2.2)

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Robitzsch, A. (2020a). Lp loss functions in invariance alignment and haberman linking with few or many groups. *Stats*, *3*(3), 246–283.

Robitzsch, A. (2020b). Robust haebara linking for many groups: Performance in the case of uniform DIF. *Psych*, *2*(3), 155–173.

Robitzsch, A. (2022). sirt: Supplementary item response theory models [Computer software manual]. Retrieved from `https://CRAN.R- project.org/package=sirt` (R package version 3.12-66)

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied psychological measurement*, *7*(2), 201–210.

Weeks, J. P. (2010). plink: IRT separate calibration linking methods. *Journal of Statistical Software*, *35*(12), 1–33. Retrieved from `http://www.jstatsoft.org/v35/i12/`