Journal of Computerized Adaptive Testing

Volume 12 Number 4

August 2025

Stochastic Curtailment: A New Approach to Improve Efficiency of Variable-Length Computerized Adaptive Tests

Ming Him Tai
Pennsylvania State University

Joseph N. DeWeese and David J. Weiss
University of Minnesota

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing
www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing

Volume 12, Number 4, August 2025

DOI 10.7333/2508-1204165

ISSN 2165-6592

Stochastic Curtailment: A New Approach to Improve Efficiency of Variable-Length Computerized Adaptive Tests

Ming Him Tai
Pennsylvania State University

Joseph N. DeWeese and David J. Weiss
University of Minnesota

Stochastic curtailment (SC) is a statistical procedure that was originally developed to enhance the efficiency of clinical trials. It has been applied to psychological testing, but to sequential mastery testing only (Finkelman, 2008, 2010). This study adapted the method to detect low-precision examinees in computerized adaptive tests (CATs) (i.e., examinees whose final standard error of measurement (FSEM) at the end of a full-length CAT could not reach the pre-specified SEM termination level). Using central limit approximations, the study developed a method to estimate the distribution of test information at maximum test length and the corresponding FSEM. The study also developed a hypothesis testing procedure to implement SC. Using monte-carlo simulations, the study found that (1) the FSEM estimation procedure performed well in the middle range of θ values but less so at extreme θ values; (2) the SC procedure had good predictive accuracy, with excellent performance on positive predictive values and good performance on true positive rates and false positive rates; (3) the potential reduction in test length was substantial. Overall, the study showed that SC is a promising procedure to identify low-precision examinees and enhance efficiency in measurement CATs. A brief guide to implementing SC is provided.

Keywords: computerized adaptive tests, low-precision CATs, stochastic curtailment, termination rules, variable- length CAT

Computerized adaptive tests (CATs), which use item response theory (IRT) and principles of artificial intelligence and machine learning to customize item selection for examinees, allow test items and test length to vary across examinees. CAT administrators commonly determine an appropriate test length for each examinee by setting a target observed standard error of measurement (SEM) level. Each additional item administered adds new information about the examinee's trait level, therefore typically reducing their SEM. The test continues administering new items until the examinee's SEM reaches the target level, upon which the test is terminated.

However, for some examinees, their SEMs can never reach the target level, even if a maximum number of items is administered. This usually happens because there is an insufficient number of informative items at and around their trait levels so that they cannot be measured precisely and reach the target SEM for termination, or because they do not respond to all items according to the underlying IRT models, resulting in conflicting information about their trait levels. These examinees are referred to as “low-precision cases.” When this happens, it is a waste of time and resources for both the examinees and the administrator. A question that has remained unanswered over half a century of CAT is: Is it possible to identify low-precision cases early in the administration of a CAT and end their test as early as possible after the test starts?

The key to identifying low-precision cases is to estimate an examinee’s SEM at a designated full or maximum test length (hereafter referred to as FSEM) when only a limited number of items have been administered. For example, if the estimated probability that an examinee’s FSEM will reach the target SEM level is only 1%, i.e., there is a 99% chance that the person is a low-precision case, then the test can be terminated early. While traditional termination rules [e.g., SEM, minimum information (MI), and others] aim to balance precision and test length, they lack a mechanism to probabilistically predict whether an examinee will ultimately achieve the target SEM. What is needed is a method to estimate, early in the test, the probability that an examinee’s final SEM (FSEM) will meet the target threshold—a capability absent in existing CAT termination rules.

This gap motivates the application of stochastic curtailment (SC), a statistical procedure originally developed for medical clinical trials (Davis & Hardy, 1994). SC evaluates whether accumulating data will likely meet a predefined outcome, enabling early termination with minimized risk. Adapting SC to CAT offers a novel solution to the longstanding question: Can we identify low-precision cases early in a CAT and terminate their tests efficiently? To date, although there have been some applications of SC to psychological testing using sequential testing procedures (Finkelman, 2008, 2010; Finkelman et al., 2011, 2012; Sie et al., 2015), there is no SC procedure applicable to the measurement of trait levels using CATs.

This research has two major goals: (1) developing and evaluating a statistical procedure to estimate FSEM during a CAT process, and (2) developing and evaluating a SC procedure based on the FSEM estimation procedure. It will not only contribute to the literature on SC and test termination in CAT, but also bring substantial benefits to those implementing CATs by saving time and resources for both examinees and administrators in low-precision cases.

Related Research

The termination rule of a variable-length CAT, i.e., under what conditions will the CAT be ended, is one of its essential elements (Weiss & Kingsbury, 1984; van der Linden & Glas, 2010; Weiss & Şahin, 2024). One of the most commonly used termination rules is the SEM rule, which states that the test proceeds until the observed SEM of the θ estimate equals or falls below a pre-specified level (Weiss & Kingsbury, 1984). Theoretically, the SEM rule ensures that every examinee can be measured with equal precision regardless of their true θ levels, as long as the item bank has sufficient information across the θ regions of interest (Dodd et al., 1989, 1993; Revuelta & Ponsoda, 1998; Wang & Wang, 2001). However, in cases when some regions of θ do not have sufficient informative items, examinees whose θ levels are in those regions might not reach the pre-specified SEM level, even if all items in the bank are exhausted. A common scenario under which low-precision cases will occur is examinees with extreme θ levels tested under an item bank with a peaked information function, in which the number of informative items at extreme θ levels is small (Choi et al., 2011).

A number of alternative termination rules have been proposed to deal with the limitations of the SEM rule, such as the MI rule (Gialluca & Weiss, 1979; Maurelli & Weiss, 1981), the predicted standard error reduction (PSER) rule (Choi et al., 2011), and the change in θ (CT) rule (Babcock & Weiss, 2012; Wang et al., 2019). Each rule has been shown to be able to reduce the needless administration of uninformative items to some extent, but each has its own limitations (Babcock & Weiss, 2012; Wang et al., 2019). The MI rule stops the test once every remaining item falls below a fixed information threshold; in banks with peaked information functions this can truncate tests prematurely, because a large bank of low-discrimination items might still, in aggregate, provide enough information to attain the target SEM had they been administered, thereby biasing scores for examinees located in the extremes of the trait continuum. The PSER rule projects how much the SEM would shrink if another item were delivered and ends the test when the expected gain is minimal; however, those projections are model-based and can be badly mis-calibrated when the current θ estimate is unstable or when the algorithm's look-ahead horizon is set too short—each scenario inflating premature terminations and widening between-person precision inequities. Finally, the CT rule monitors successive θ estimates and terminates once their absolute difference drops below a pre-specified threshold; because it ignores the SEM, it can declare convergence even when information is sparse (yielding imprecise scores), or it can prolong testing for examinees near item-bank information peaks where θ estimates oscillate minutely but SEM is already satisfactory; moreover, its convergence constant is typically chosen ad hoc, so its operating characteristics vary unpredictably across banks and θ regions.

Researchers have also proposed compound rules that combine the SEM rule and another rule to yield the advantages of both, such as the SEM+MI rule (Babcock & Weiss, 2012) and the SEM + change-in-θ rule (Wang et al., 2019). However, there are still unresolved issues using these compound rules. First, compound rules might still lead to premature termination; for example, the SEM+MI rule states that a CAT will be terminated if either the observed SEM drops below the pre-specified level under the SEM rule or all non-administered items have less information than some specified minimal amount of information (Babcock & Weiss, 2012). However, for a large item bank with a peaked information function, there might be an abundant number of low-informative items such that the pre-specified SEM is attainable across the entire range of θ. But if a large number of these items carry less information than the minimal information stated in the MI rule, the CAT will be terminated without administering these items, which would have increased the measurement precision to the target level had they been administered. An example is the K-12 reading achievement item bank studied by Lee (2019). It has 1,089 items, but the average of their discrimination parameters is only .88 (with D = 1.7), meaning that many of them provide relatively small amounts of information.

Moreover, to change the termination rule from the SEM rule to another rule during a CAT, therefore sacrificing precision for efficiency, might not be aligned with the goals of the test. For example, in educational testing, measuring each examinee with equal precision (i.e., equiprecise measurement) is essential to ensure fairness (Weiss & Kingsbury, 1984); in health outcomes research, an SEM stopping rule has been used because there is a need to measure individuals with severe medical problems accurately (Ware et al., 2000, 2003, 2005).

Therefore, switching to a secondary termination rule or a compound rule does not fully resolve the issue of low precision. During a CAT process, whenever it is doubtful whether an examinee can attain the pre-specified SEM, it will be extremely helpful if the FSEM of the examinee can be estimated. If the FSEM can be predicted to reach the pre-specified SEM level, the CAT process can proceed as planned (and FSEM evaluated after each item) and potentially be terminated by the SEM rule; otherwise, the administrator might decide to terminate the CAT, change the target SEM level, or switch to a different termination rule, depending on the measurement objectives of the test. Moreover, in the CAT termination literature to date, none of these termination methods, either single or compound, estimates the FSEM or evaluates how the method functions relative to the actual SEM during the test. In order to estimate FSEM, the statistical procedure of SC can be applied.

SC procedures were initially developed for phase II clinical trials in drug evaluation (Halperin et al., 1982; Davis & Hardy, 1994; Ayanlowo & Redden, 2007; Law et al., 2020, 2022) to predict if adding more subjects to the trial has the potential to result in a defined positive outcome (e.g., improvement in symptoms for a specified proportion of patients); if the prediction is negative the trial is stopped, and if positive it is continued. SC has been adapted to psychological testing since the 2000s, but only to classification tests such as the sequential mastery test (SMT; Finkelman, 2008, 2010), in which an examinee receives a binary classification decision (e.g., master/non-master). To reduce test length, the truncated sequential probability ratio test (TSPRT; Finkelman, 2008, 2010) can be applied to determine whether sufficient information has accrued about the examinee's trait level to allow a classification decision be made. Using monte-carlo simulation designs, Finkelman (2008, 2010) applied SC and showed that it considerably reduced the average test length of a SMT with only a slight decrease in accuracy. Using a post-hoc simulation design with real response data, the procedure has also been shown to reduce the average number of questions administered with a minimal loss of classification accuracy in the Medicare Health Outcomes Survey (Finkelman et al., 2011) and the Center for Epidemiologic Studies Depression scale (Finkelman et al., 2012).

However, most psychological tests measure constructs along a trait continuum, such as a test that measures a student's math aptitude, a patient's depression level, or a job applicant's cognitive ability. CATs applied to these tests are referred to as measurement CATs in comparison to classification CATs discussed above. In these tests, the SC procedures developed for SMTs do not apply because the purpose of the test is to obtain a point estimate of θ. However, it is meaningful to specify hypotheses about the FSEM and develop a procedure to test hypotheses focusing on measurement precision. This allows the CAT algorithm to monitor whether an examinee can attain the pre-specified SEM level during the test and take the appropriate action. To date, no work has been done to develop such a procedure. This research is designed to fill this gap.

Method

Development

To perform SC, it is essential to estimate FSEM when only a limited number of items have been administered. Development of the method proceeds in two steps: (1) derive the probability distribution of FSEM based on the central limit theorem (CLT), which is a statistical technique commonly used in sequential analysis (Finkelman, 2008); and (2) develop a SC rule based on the probability distribution of FSEM derived in the previous step. The rule will be a hypothesis testing procedure to determine if SC should be applied for an examinee.

Probability Distribution of FSEM

Assume a CAT process in which the items have dichotomously or polytomously scored responses. The maximum test length is N. After administering k items, given the uncertainty in the selection of the remaining $N - k$ items and the estimate of θ, FSEM can be treated as a random variable that follows a probability distribution.

Denote FSEM as $sem(\hat{\theta}_N|\boldsymbol{u}_k)$, where $\boldsymbol{u}_k$ denotes the response vector for the k administered items and $\hat{\theta}_N$ denotes an estimator of the examinee's θ level at the pre-specified maximum test length. Importantly, the asymptotic variance of $\hat{\theta}_N$ equals the reciprocal of its Fisher information under maximum likelihood estimation (MLE). Therefore, there is a one-to-one relationship between $sem(\hat{\theta}_N|\boldsymbol{u}_k)$ and the observed information (van der Linden & Glas, 2010, p.16) of the test, denoted as $J(\hat{\theta}_N|\boldsymbol{u}_k)$:

$$J(\hat{\theta}_N|\boldsymbol{u}_k) = \frac{1}{Var(\hat{\theta}_N|\boldsymbol{u}_k)} = \frac{1}{\left[sem(\hat{\theta}_N|\boldsymbol{u}_k)\right]^2}, \tag{1}$$

where the observed information of a test or an item is defined as the negative of the second-order derivative of its likelihood function conditional on the θ estimate. More technical details can be found in Appendix A.

The following discussion focuses on $J(\hat{\theta}_N|\boldsymbol{u}_k)$ instead of $sem(\hat{\theta}_N|\boldsymbol{u}_k)$ because $J(\hat{\theta}_N|\boldsymbol{u}_k)$ can be expressed as the sum of individual item information, assuming local independence among items. This relationship enables the application of CLT-based approximations by taking advantage of the additivity of individual item information. Assuming $N - k$ is sufficiently large, the CLT implies that

$$J(\hat{\theta}_N|\boldsymbol{u}_k) \sim N(\mu, \sigma^2)\,, \tag{2}$$

where $\mu = E\left[J(\hat{\theta}_N|\boldsymbol{u}_k)\right]$, $\sigma^2 = Var\left[J(\hat{\theta}_N|\boldsymbol{u}_k)\right]$. Equation 2 says that (1) $J(\hat{\theta}_N|\boldsymbol{u}_k)$ approximately follows a normal distribution; and (2) the mean and variance of this normal distribution can be approximated by the empirical mean and variance of the observed information of the test, which can be calculated using the formulas in Appendix B.

There are two obstacles to estimating the empirical information's mean and variance: First, after k administered items, the remaining $N - k$ items are unknown, so an item projection method is needed to select $N - k$ items from the item bank; second, $\hat{\theta}_N$ is unknown, so it needs to be estimated. There are three approaches to resolve these issues, namely maximum information (MI), maximum posterior-weighted information with uniform prior (MPWI-U), and maximum posterior-weighted information with a normal prior (MPWI-N). The details are described in Appendix C.

The SC Rule

After deriving the probability distribution of $J(\hat{\theta}_N|\boldsymbol{u}_k)$, a null-hypothesis significance test can be conducted after each CAT item to determine whether SC should be triggered. The null and alternative hypotheses are,

$$H_0 : \tau \leq \tau_0 \;\; vs. \;\; H_a : \tau > \tau_0 \, , \tag{3}$$

where τ is the true FSEM of the examinee, and τ_0 is the pre-specified SEM threshold that will terminate the test. Because of the inverse relationship between information and SEM, an alternative way to state the hypotheses is:

$$H_0 : J \geq J_0 \;\; vs. \;\; H_a : J < J_0 \, , \tag{4}$$

where J is the true observed information of the examinee, J_0 is the amount of information corresponding to τ_0, i.e., $J_0 = 1/\tau_0^2$. The null hypothesis states that the FSEM is equal to or below the threshold; these examinees are referred to as *high-precision cases*. The alternative hypothesis states that the FSEM is greater than the threshold; they are referred to as *low-precision cases*. Based on the distribution of FSEM derived in the previous section, the test statistic can be constructed as

$$Test\ statistic = \frac{E\left[J\left(\hat{\theta}_N | \boldsymbol{u}_k\right)\right] - \frac{1}{\tau_0^2}}{\sqrt{Var\left[J\left(\hat{\theta}_N | \boldsymbol{u}_k\right)\right]}} \tag{5}$$

and compared to $z_{1-\alpha}$, which is the critical value at the $(1-\alpha)^{th}$ quantile of the standard normal distribution. If the test statistic is less than $z_{1-\alpha}$, then do not reject the null (the examinee is a high-precision case) and do not perform SC; if the test statistic is equal to or larger than $z_{1-\alpha}$, then reject the null (the examinee is a low-precision case) and perform SC.

Evaluation

Method

The FSEM estimation procedure and the SC rule were evaluated with simulated response data and real data, based on its performance under a variety of conditions assumed to affect its performance. The following design factors were considered in the monte-carlo simulations:

1. *Item bank characteristics*, which is essential to the operation of CAT (Weiss & Şahin, 2024);
2. *True θ level*. Because item bank information usually varies across the θ continuum, the performance of the FSEM estimation procedure and the SC rule for examinees with different θ levels were expected to be different;
3. *Maximum test length (N) and target FSEM level (τ).* Longer tests result in smaller FSEMs, so N and τ were jointly investigated;
4. *Item projection method,* i.e., the three methods for projecting items for the remaining CAT; and
5. *Starting point of applying the SC rule,* i.e., after how many administered items should a CAT start to apply the SC rule, which is an important issue in its implementation.

Item bank characteristics. Three item banks were evaluated. Two banks of dichotomously scored items were simulated, each with 200 items. Their item parameters followed the distributions in Table 1. Item parameters were for the three-parameter logistic model with D = 1.7. The third bank was an item bank with ordinal polytomous items from a real dataset. It measured applied cognition based on patient reports obtained in a hospital setting (Wang et al., 2022) using the graded response model (Samejima, 1968). The distribution of the item parameters for the dichotomous banks are presented in Table 1 and for the polytomous bank are presented in Table 2.

Table 1. Item characteristics of the two dichotomous item banks

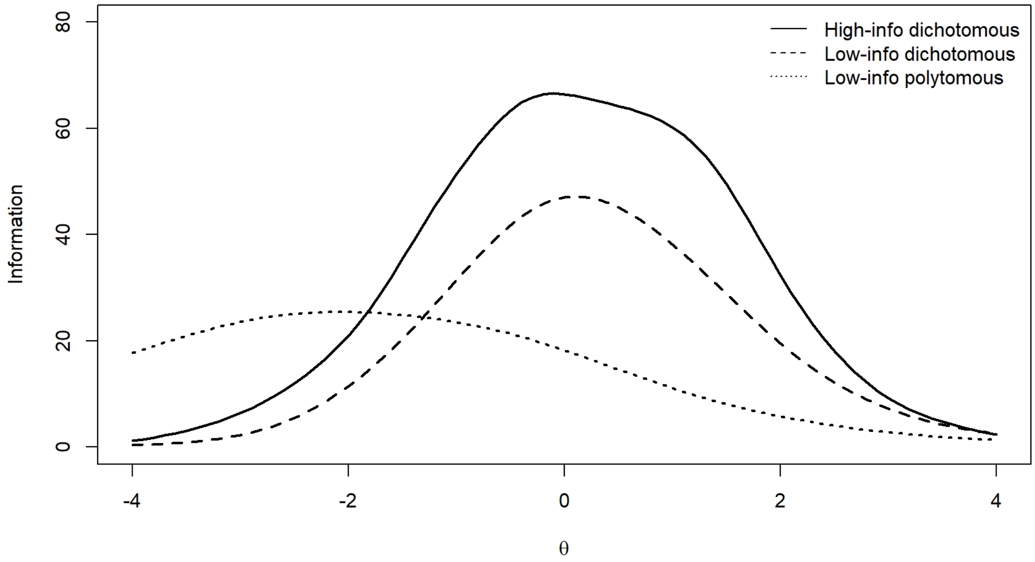
Item Bank	Discrimination (a_i)	Difficulty (b_i)	Guessing (c_i)
High-information	$\mathrm{N}(1.25, .25^2)$	$\mathrm{N}(0, 1.2^2)$	.2
Low-information	$\mathrm{N}(.8, .25^2)$	$\mathrm{N}(0, .8^2)$	.2

Table 2. Item parameters of the low-information real polytomous item bank

Statistic	Discrimination (a_i)	Boundary 1 (b_{i1})	Boundary 2 (b_{i2})	Boundary 3 (b_{i3})
Mean	.81	-4.13	-2.32	-.73
S.D.	.18	.85	.70	.83

The distributions of the item parameters for the two simulated dichotomous item banks were determined in reference to real item banks. The high-information bank was similar to a bank that measures K-12 math ability (Phadke, 2017) and the low-information bank was similar to a bank that measures K-12 reading ability (Lee, 2019). Some minor adjustments on the parameters were made to make their bank information functions more differentiated, as depicted in Figure 1.

Figure 1. Information functions of the three item banks



This study did not investigate item banks with equal measurement precision across the θ continuum (Kim-Kang & Weiss, 2008) because this study focused on low-precision caused by an insufficient number of informative items in regions of the bank. Other causes of low-precision, such as examinees not responding to items according to the underlying IRT models, which affects observed information, were not investigated in the simulations. Therefore, only item banks with an inverted-U shape information pattern, which might cause insufficient test information in regions of θ, were generated for the dichotomous banks.

True θ level. For the two dichotomous banks, five true θ levels were investigated: $\theta = \{-2, -1\ ,0, 1,2\}$ (Lee, 2015; Phadke, 2017). There were 1,000 simulees at each θ level. Given the peaked shape of the bank information functions, these values were chosen to investigate how the centeredness or extremeness of true θ levels, corresponding to different levels of bank information, affected the performance of the FSEM estimation procedure and the SC rule. Note that the information functions of the two dichotomous banks peaked at around $\theta = 0$, therefore the θ levels were chosen to be symmetric about 0. For the polytomous bank, seven true θ levels were investigated: $\theta = \{-4, -3 - 2, -1, 0, 1, 2\}$. This bank peaked at around $\theta = -2$, so θ levels from –4 to 0 were chosen. In addition, $\theta = 1$ and $\theta = 2$ were chosen to allow comparison with the dichotomous banks at those trait levels.

N and τ. Using the three item banks, sets of preliminary simulations were run to determine the relationship between N and FSEM and the appropriate values of τ; These simulations are described in Appendix D along with their results. Table 3 displays the τ values selected for each of the item banks and test lengths.

Table 3. Termination SEMs (τ) for each item bank and test length (N)

Item Bank	$N = 20$	$N = 30$	$N = 40$
High-information dichotomous	.26	.23	.20
Low-information dichotomous	.33	.30	.27
	$N = 15$	$N = 20$	$N = 25$
Low-information polytomous	.43	.40	.37

Item projection method. The uncertainty in selecting the $N - k$ items is a major source of uncertainty in estimating FSEMs. The three item projection approaches discussed in the previous section, namely MI, MPWI-N, MPWI-U, were evaluated.

Starting point of applying SC rule. The simulation process began to evaluate the SC rule after administering 5 items, because previous work showed that the performance of SC is unsatisfactory before 5 items for polytomous item banks (DeWeese & Weiss, 2023). The evaluation process continued after every item until the $(N - 1)^{th}$ item, which is the last possible item to apply the SC rule. In other words, denoting the number of administered items as k, the SC rule was evaluated at $k = 5, 6, \ldots, N - 1$, which totals $N - 5$ evaluation points. The performance of the SC rule at each value of k against the evaluation criteria, which are discussed below, determined the optimal value of k to apply the SC rule under an item bank and the other manipulated variables.

Summary. For each dichotomous item bank, the number of simulation conditions given N equaled the product of 5 (true θ levels) × 3 (*N and τ values*) × (*N* – 5) (*evaluation points*). For the polytomous bank, the number of simulation conditions given N was 7 (true θ levels) × 3 (*N and*

τ values) × (*N* – 5) (*evaluation points*). Table 4 lists the numbers of simulation conditions. The total number of simulation conditions for this study was $1{,}125 \times 2 + 945 = 3{,}195$.

Table 4. Number of simulation conditions for the three banks

Item Bank	*N*=20	*N*=30	*N*=40	*Total*
High-information dichotomous	225	375	525	1,125
Low-information dichotomous	225	375	525	1,125
	N=15	*N*=20	*N*=25	*Total*
Polytomous	210	315	420	945

Evaluation of the Method

Under each simulation condition (a combination of an item bank, a fixed θ level, a set of N and τ values, and an evaluation point k), the following process was randomly replicated 1,000 times:

1. Administer the first 5 items. The starting θ was 0. θ was estimated with MLE. If responses were uniform, then maximum a priori estimation with a standard normal prior was used until a mixed response pattern was obtained. Items were chosen to maximize the expected Fisher information conditional on the current θ estimate.
2. Obtain the probability distribution of test information (TI) after administering k items, where $k = 5, 6, \dots, N-1$. Then apply the SC rule. If the null hypothesis is rejected at the 5% level, the examinee will be predicted as a low-precision case. Otherwise, it will be predicted as a high-precision case. In addition, estimate FSEM using the formula $FSEM = 1/\sqrt{\text{TI}}$, where the value of TI is the mean of its distribution.
3. Administer the remaining $N-k$ items and estimate θ using MLE. Obtain the actual FSEM. If the actual FSEM is greater than the threshold level τ, the examinee is an actual low-precision case. Otherwise, it is an actual high-precision case.

A graphical illustration of the procedures described above is presented in Figure 2.

Evaluation Criteria

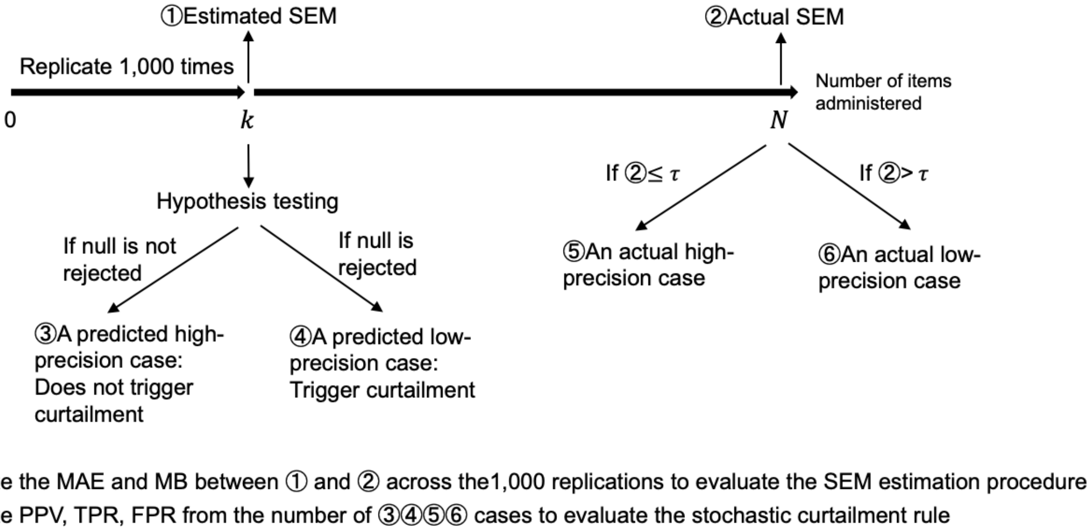
FSEM estimation procedure. To evaluate the performance of the FSEM estimation procedure, mean absolute error (MAE) and mean bias (MB) between the estimated FSEM (from Step 3 above) and the actual FSEM (from Step 2 above) were computed. They are defined as

$$MAE_{FSEM} = \frac{\sum_{i=1}^{n} |FSEM_{est \cdot i} - FSEM_{actual \cdot i}|}{n} \quad (6)$$

$$MB_{FSEM} = \frac{\sum_{i=1}^{n} (FSEM_{est \cdot i} - FSEM_{actual \cdot i})}{n} \quad (7)$$

where $FSEM_{est \cdot i}$ is the estimated FSEM in the i^{th} replication, $FSEM_{actual \cdot i}$ is the actual FSEM in the i^{th} replication, and n is 1,000. MAE indicates the magnitude of the estimation error, or how much the estimated FSEM deviates from the actual FSEM, on average. MB indicates whether estimated FSEM is larger or smaller than actual FSEM, on average. A positive MB means that the estimated FSEM is larger than the actual FSEM, on average, whereas a negative MB means that estimated FSEM is smaller than actual FSEM, on average.

Figure 2. Evaluation process of the FSEM estimation procedure and the SC rule


①Estimated SEM
②Actual SEM
Replicate 1,000 times
Number of items administered
0
k
N
Hypothesis testing
If null is not rejected
If null is rejected
③A predicted high-precision case: Does not trigger curtailment
④A predicted low-precision case: Trigger curtailment
If ②≤ τ
If ②> τ
⑤An actual high-precision case
⑥An actual low-precision case

- Compute the MAE and MB between ① and ② across the1,000 replications to evaluate the SEM estimation procedure
- Compute PPV, TPR, FPR from the number of ③④⑤⑥ cases to evaluate the stochastic curtailment rule

SC rule. To evaluate the performance of the SC rule, positive predictive value (PPV), true positive rate (TPR), and false positive rate (FPR) were examined. They are defined in Table 5.

Table 5. Hypothesis testing results and metrics for evaluating the SC rule

		Predicted state		
		High-precision (negative)	Low-precision (positive)	
True state	High-precision (negative)	Correct (true negative)	Type I error (false positive)	$FPR = FP/(FP + TN)$
	Low-precision (positive)	Type II error (false negative)	Correct (true positive)	$TPR = TP/(FN + TP)$
			$PPV = TP/(FP + TP)$	

High-precision was defined as negative and low-precision was defined as positive. The most important criterion was designated as PPV, which is the proportion of procedure-predicted low-precision cases that actually were low-precision. It answered the question “When the procedure predicts that the examinee is a low-precision case, how confident are we in this prediction?” PPV was important because whenever the procedure predicts a low-precision case, the test will be curtailed. A low PPV means a large proportion of the predicted low-precision cases, which actually were high-precision cases, will be terminated prematurely. From a practical perspective, this is highly undesirable, because these examinees could have reached the termination SEM.

In many simulation studies, recovery of true value is evaluated by TPR, also known as power (Wang et al., 2021; 2019). In this context, TPR was the proportion of low-precision cases in the population that can be correctly identified by the procedure. While a high TPR is desirable, a low TPR does not have dire consequences because if a low-precision case is mistakenly predicted as a high-precision case, the test just continues. The examinee will take unnecessary items, but it is less detrimental than the scenario that a high-precision examinee is mistakenly predicted as a low-precision case and therefore terminated prematurely, which is captured by a low PPV. Nevertheless, TPR is still an important measure of the efficiency of the SC procedure, therefore it was considered as a secondary criterion.

FPR was defined as the proportion of high-precision cases in the population that were mistakenly predicted as low-precision cases by the procedure. These cases will be mistakenly curtailed, resulting in premature termination of the test, which is undesirable. However, at extreme θ values, which were of primary interest in this study, only a small portion of the examinees were high-precision cases. Therefore, even when the FPR was high, the actual number of false positive cases will still be small. Therefore, FPR was used as another secondary criterion.

Test length. In addition, the reductions in average test length (ATL) were computed to determine the effectiveness of the SC procedure. Specifically, for each simulee, a CAT without SC was simulated using the specified value of τ to determine the actual test length. Then, a CAT with SC was simulated, and the number of items administered before SC was triggered was the test length under SC. The difference between the actual test length and the test length under SC, averaged across the number of simulees, was the reduction in ATL.

Software

All simulations were conducted using R statistical software (R Core Team, 2024).

Results

FSEM Estimation

Results are presented below for the 20-item CATs because that test length was common across the two item formats. Detailed results for other test lengths are in Appendix E and Appendix F.

MAE. The MAE results in Figure 3 and Table 6 show several consistent patterns for MAE with 20-item CATs: (1) For all θ values and item projection methods, MAE decreased as the test proceeded; (2) the MAEs at the center of the θ continuum were lower than at the extreme; (3) the MAEs under the low-information bank (Figure 3b) were generally higher than under the high-information bank (Figure 3a); (4) the three item projection methods generally had similar performance; in the few conditions where they differed it was primarily at the beginning of the tests. Similar patterns were observed for tests with other maximum test lengths; see Appendix Figures E4, E5, and E6.

MB. Generally, Figure 4 and Table 7 show that MB remained stable under most conditions, with slight decreases in the early stages of the CATs. In addition, none of the three item projection methods showed consistent directional bias across the three item banks. MB was larger at extreme

θ values. Among the three item projection methods, MPWI-N had the largest bias particularly for longer tests (Figures E8 and E9) for longer CATs drawn from low information banks.

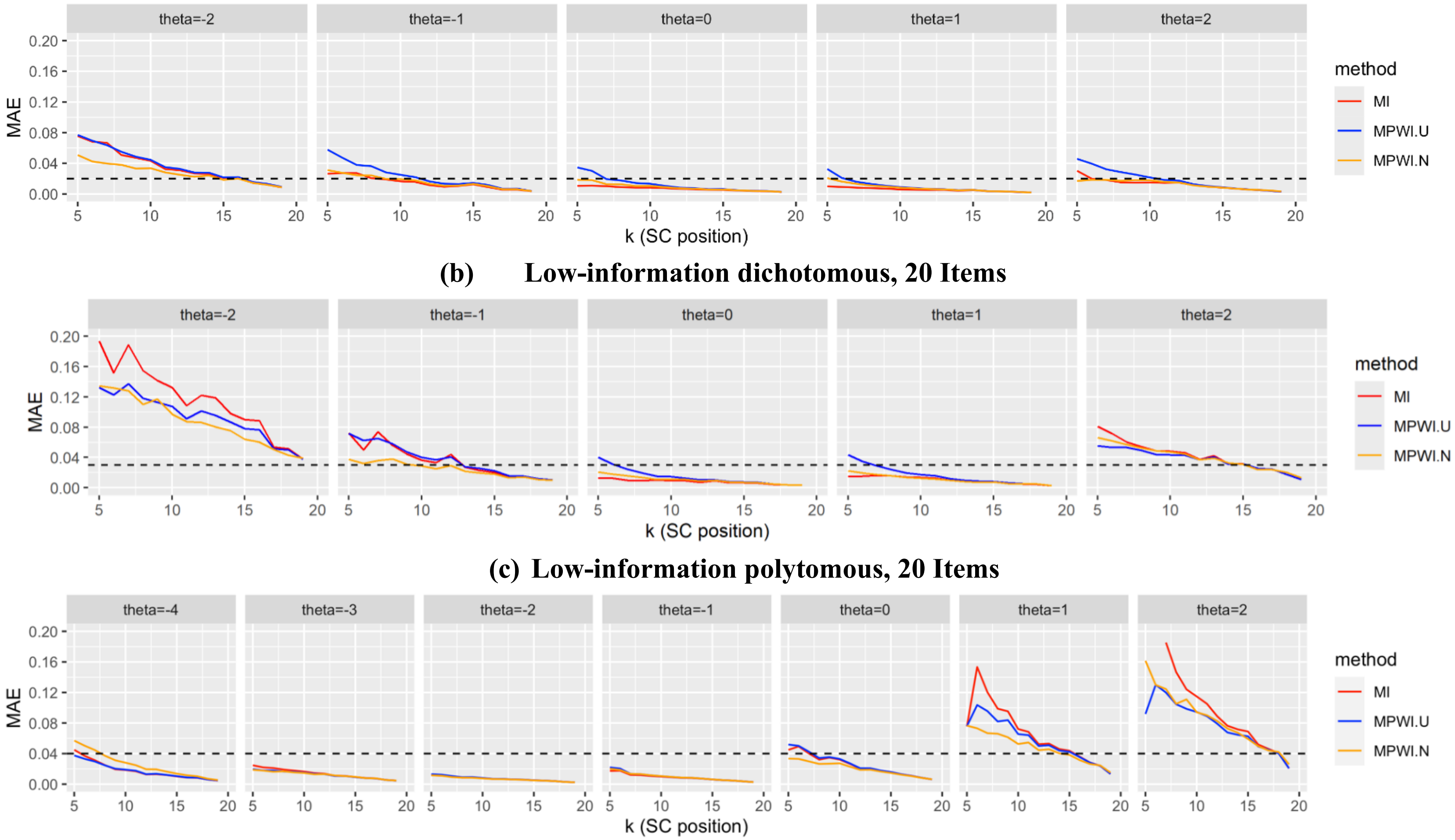
SC Procedure

PPV. PPV results are presented for extreme θ values only, because extreme θ values were the primary focus of the study. Given the high proportion of true low-precision cases at these θ values, a PPV of 80% was set as the benchmark for the procedure, a value that is recommended by some medical researchers (e.g., Winters et al., 2016) for PPV. In general, Figure 5 and Table 8 show that the PPVs were above 80% under most circumstances. Importantly, PPVs were higher when bank information was low. Comparing the low-information dichotomous bank (Figure 5b) to the high-information dichotomous bank (Figure 5a), the PPVs were almost uniformly higher. Under the polytomous bank (Figure 5c), PPVs were almost uniformly 100% under the most extreme θ values of $\theta = 1$ and 2. Under less extreme values of −4 and 0, PPVs were still high, but they were lower and fell below 80% for the first few items when $\theta = 0$. The three item projection methods had similar performance. Similar patterns were observed for other maximum test lengths except for some minor differences; at $\theta = 2$ under both dichotomous banks, the PPVs were slightly lower for the 30-item tests than the 20-item and 40-item tests (Figures E10 and E11), and at $\theta = 0$ under the polytomous bank, the PPVs were slightly higher for the 15-item test and slightly lower for the 25-item test (Figure E12).

TPR. Optimal power is widely accepted as 0.8 based on Cohen's (1992) recommendations to balance high power with demands on the researcher to recruit enough participants. TPR's performance was determined by the complex interactions among item bank, θ value, SC position, and item projection method. Several patterns were observed (Figure 6 and Table 9): (1) TPRs under low-information scenarios were higher than high-information scenarios; (2) MPWI-U performed better than the other two methods, especially at the beginning of the test under the two dichotomous item banks. In particular, the TPRs were above 80% under the low-information dichotomous bank and fluctuated between 60% and 90% under the high-information dichotomous bank; (3) at $\theta = 1$ and $\theta = 2$ under the low-information polytomous bank, TPRs were close to 100%. Similar patterns were observed for other maximum test lengths except for small differences. At $\theta = 0$ under the polytomous bank, TPRs decreased slightly as maximum test length increased (Figures E13–E15).

FPR. As shown in Figure 7 and Table 10, the patterns of FPR were highly unstable, primarily due to the low number of high-precision cases at extreme θ values. In particular, when $\theta = 2$ under the polytomous bank, there were no high-precision cases and FPRs were non-existent. Generally, as the test proceeded, FPRs decreased. In addition, MPWI-U performed worse than the other two methods, especially under the dichotomous item banks. Similar patterns were observed for other maximum test lengths (Figures E16–E18).

Figure 3. MAE of the FSEM estimation procedure for 20-item CATs



(a) High-information dichotomous, 20 Items

(b) Low-information dichotomous, 20 Items

(c) Low-information polytomous, 20 Items

Note: The black horizontal dashed lines represent the performance benchmarks, which were set at .02, .03, and .04 for the three banks, respectively. They equaled roughly 10% of the τ values in Table 3.

Table 6. MAE of the FSEM estimation procedure for 20-item CAT

(a) High-information dichotomous, 20 Items

Method	θ	SC point														
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
MI	-2	.076	.068	.067	.051	.047	.043	.032	.031	.027	.026	.021	.022	.014	.012	.009
	-1	.026	.027	.027	.021	.019	.016	.016	.011	.009	.010	.012	.009	.006	.006	.004
	0	.011	.011	.010	.009	.008	.008	.008	.007	.006	.005	.005	.004	.004	.004	.003
	1	.010	.009	.008	.008	.007	.006	.006	.005	.005	.004	.005	.004	.003	.003	.002
	2	.030	.020	.018	.015	.015	.015	.014	.015	.012	.010	.008	.007	.005	.005	.003
MPWI-N	-2	.051	.042	.040	.038	.033	.034	.028	.025	.022	.024	.018	.020	.014	.012	.009
	-1	.031	.028	.024	.024	.020	.019	.018	.013	.011	.011	.013	.010	.006	.006	.004
	0	.019	.018	.013	.012	.010	.011	.009	.007	.007	.006	.006	.005	.004	.004	.003
	1	.021	.016	.013	.011	.010	.008	.007	.006	.006	.005	.005	.004	.003	.003	.002
	2	.017	.018	.018	.017	.018	.018	.016	.015	.011	.009	.008	.007	.005	.005	.003
MPWI-U	-2	.077	.069	.064	.055	.049	.045	.035	.033	.028	.028	.022	.022	.016	.013	.009
	-1	.058	.047	.038	.036	.028	.025	.022	.016	.013	.013	.014	.012	.007	.007	.004
	0	.035	.030	.019	.018	.014	.013	.011	.008	.008	.006	.006	.005	.004	.004	.003
	1	.033	.021	.016	.013	.011	.009	.008	.007	.006	.005	.005	.004	.003	.003	.002
	2	.046	.040	.032	.029	.025	.021	.018	.017	.013	.010	.009	.007	.006	.005	.003

(b) Low-information dichotomous, 20 Items

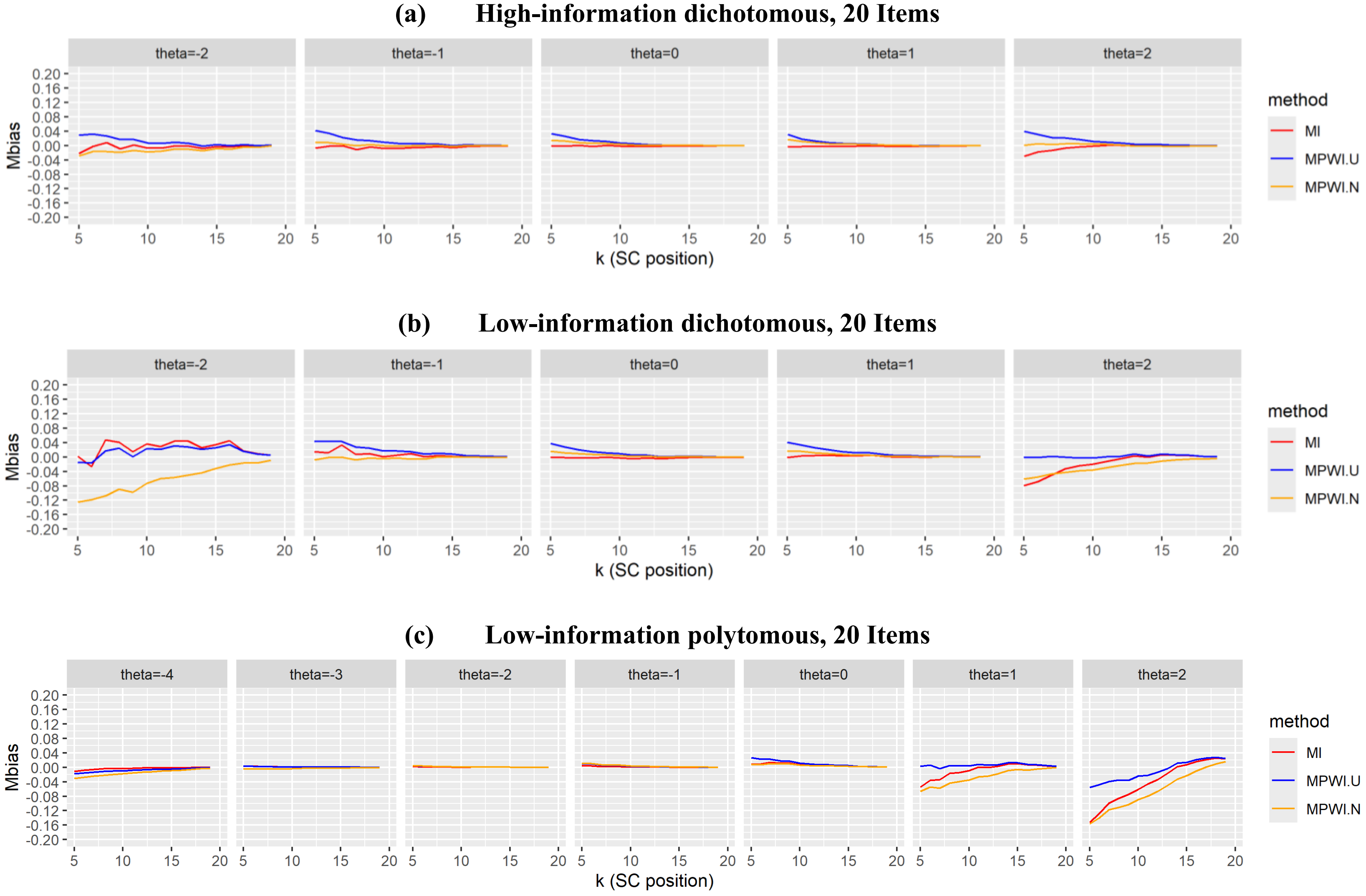
Method	θ	SC point														
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
MI	-2	.193	.152	.189	.155	.142	.133	.109	.122	.119	.098	.090	.089	.054	.052	.038
	-1	.072	.050	.074	.056	.045	.036	.033	.044	.027	.023	.020	.014	.015	.012	.010
	0	.013	.013	.010	.009	.010	.009	.010	.007	.009	.007	.007	.006	.005	.004	.003
	1	.015	.015	.016	.016	.014	.014	.013	.010	.008	.008	.008	.006	.005	.005	.003
	2	.081	.072	.060	.054	.049	.048	.046	.038	.043	.032	.032	.026	.024	.017	.011
MPWI-N	-2	.134	.132	.128	.110	.117	.097	.087	.087	.080	.076	.064	.061	.051	.043	.039
	-1	.038	.032	.036	.038	.031	.029	.025	.029	.022	.020	.018	.013	.014	.011	.010
	0	.021	.018	.016	.013	.011	.012	.011	.009	.010	.008	.007	.007	.005	.004	.003
	1	.022	.020	.017	.016	.014	.013	.012	.010	.008	.007	.008	.006	.005	.005	.003
	2	.066	.062	.057	.052	.049	.047	.045	.037	.039	.033	.031	.024	.024	.020	.013
MPWI-U	-2	.132	.122	.137	.118	.113	.107	.091	.102	.096	.087	.078	.077	.051	.050	.038
	-1	.072	.062	.065	.058	.047	.040	.037	.041	.028	.025	.022	.016	.015	.012	.010

0	.040	.031	.024	.019	.015	.015	.013	.010	.010	.008	.008	.007	.005	.004	.003
1	.044	.035	.029	.024	.020	.017	.016	.012	.010	.009	.008	.007	.006	.005	.003
2	.055	.053	.053	.050	.044	.044	.043	.037	.041	.031	.032	.026	.024	.017	.011

(c) Low-information polytomous, 20 Items

Method	θ	SC point														
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
MI	-4.0	.045	.037	.031	.025	.019	.018	.017	.013	.013	.012	.010	.009	.008	.006	.005
	-3.0	.025	.022	.021	.019	.018	.016	.014	.014	.011	.011	.009	.008	.007	.006	.005
	-2.0	.012	.011	.010	.009	.009	.008	.007	.007	.006	.006	.005	.005	.004	.003	.002
	-1.0	.017	.018	.012	.012	.011	.010	.009	.008	.008	.007	.006	.005	.005	.004	.003
	.0	.045	.049	.040	.032	.035	.032	.027	.020	.021	.018	.016	.013	.011	.008	.006
	1.0	.076	.153	.120	.099	.095	.072	.068	.052	.053	.046	.044	.036	.028	.024	.013
	2.0	.143	.255	.185	.147	.124	.115	.105	.089	.076	.072	.069	.052	.046	.041	.021
MPWI-N	-4.0	.057	.050	.045	.037	.031	.028	.025	.020	.019	.016	.014	.012	.010	.007	.005
	-3.0	.019	.018	.016	.017	.015	.015	.013	.013	.011	.011	.010	.008	.008	.006	.005
	-2.0	.012	.011	.009	.009	.008	.008	.007	.007	.006	.006	.005	.005	.004	.003	.002
	-1.0	.020	.019	.013	.013	.012	.011	.009	.009	.008	.007	.006	.005	.005	.004	.003
	.0	.033	.033	.030	.026	.027	.027	.023	.019	.019	.017	.015	.012	.011	.008	.006
	1.0	.077	.073	.067	.066	.061	.053	.054	.044	.046	.040	.038	.031	.026	.025	.015
	2.0	.162	.130	.124	.105	.111	.094	.090	.083	.073	.067	.059	.050	.044	.042	.026
MPWI-U	-4.0	.038	.033	.030	.025	.020	.019	.018	.013	.014	.012	.010	.009	.009	.006	.005
	-3.0	.020	.018	.018	.017	.016	.015	.014	.013	.011	.011	.009	.008	.007	.006	.005
	-2.0	.013	.012	.011	.009	.009	.008	.007	.007	.006	.006	.005	.005	.004	.003	.002
	-1.0	.022	.020	.014	.013	.012	.010	.009	.008	.008	.007	.006	.005	.005	.004	.003
	.0	.052	.050	.042	.034	.035	.033	.027	.021	.021	.018	.016	.013	.011	.008	.006
	1.0	.076	.104	.096	.082	.084	.065	.064	.050	.051	.044	.042	.035	.028	.024	.013
	2.0	.092	.130	.120	.105	.099	.095	.089	.079	.068	.065	.063	.049	.045	.040	.021

Figure 4. MB of the FSEM estimation procedure for 20-item CATs



(a) High-information dichotomous, 20 Items

(b) Low-information dichotomous, 20 Items

(c) Low-information polytomous, 20 Items

Table 7. MB of the FSEM estimation procedure for 20-item CATs

(a) High-information dichotomous, 20 Items

Method	θ	SC point														
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
MI	-2	-.022	-.002	.008	-.009	.002	-.007	-.007	.000	-.002	-.008	-.002	-.003	.000	-.004	.000
	-1	-.007	-.002	-.001	-.011	-.004	-.008	-.008	-.006	-.004	-.003	-.006	-.002	-.001	-.002	-.001
	0	-.002	-.001	.001	-.002	.000	-.002	-.003	-.002	-.002	-.002	-.001	-.001	-.001	.000	.000
	1	-.004	-.003	-.002	-.003	-.002	-.002	-.002	-.002	-.002	-.002	-.002	-.001	-.001	-.001	.000
	2	-.030	-.017	-.013	-.006	-.003	-.002	.001	.001	.000	.000	.000	-.001	.000	-.001	-.001
MPWI-N	-2	-.028	-.017	-.017	-.019	-.013	-.018	-.015	-.009	-.010	-.014	-.008	-.010	-.004	-.005	-.001
	-1	.009	.009	.004	.000	.003	.000	-.001	.000	.001	.000	-.003	.001	.000	.000	.000
	0	.015	.013	.009	.006	.006	.004	.002	.001	.001	.001	.000	.000	.000	.001	.000
	1	.017	.012	.008	.006	.004	.004	.003	.001	.001	.001	-.001	.000	.000	.000	.000
	2	.000	.005	.003	.005	.005	.004	.004	.001	-.001	-.001	-.002	-.002	-.002	-.002	-.001
MPWI-U	-2	.029	.032	.027	.018	.017	.007	.006	.009	.006	-.001	.003	.000	.003	-.001	.002
	-1	.042	.034	.023	.016	.014	.009	.006	.005	.005	.004	.000	.003	.001	.001	.000
	0	.033	.026	.017	.013	.011	.008	.005	.003	.002	.001	.001	.001	.000	.001	.000
	1	.031	.018	.012	.009	.006	.005	.004	.002	.001	.001	.000	.000	.000	.000	.000
	2	.040	.032	.022	.021	.017	.012	.010	.007	.004	.003	.002	.001	.001	.000	.000

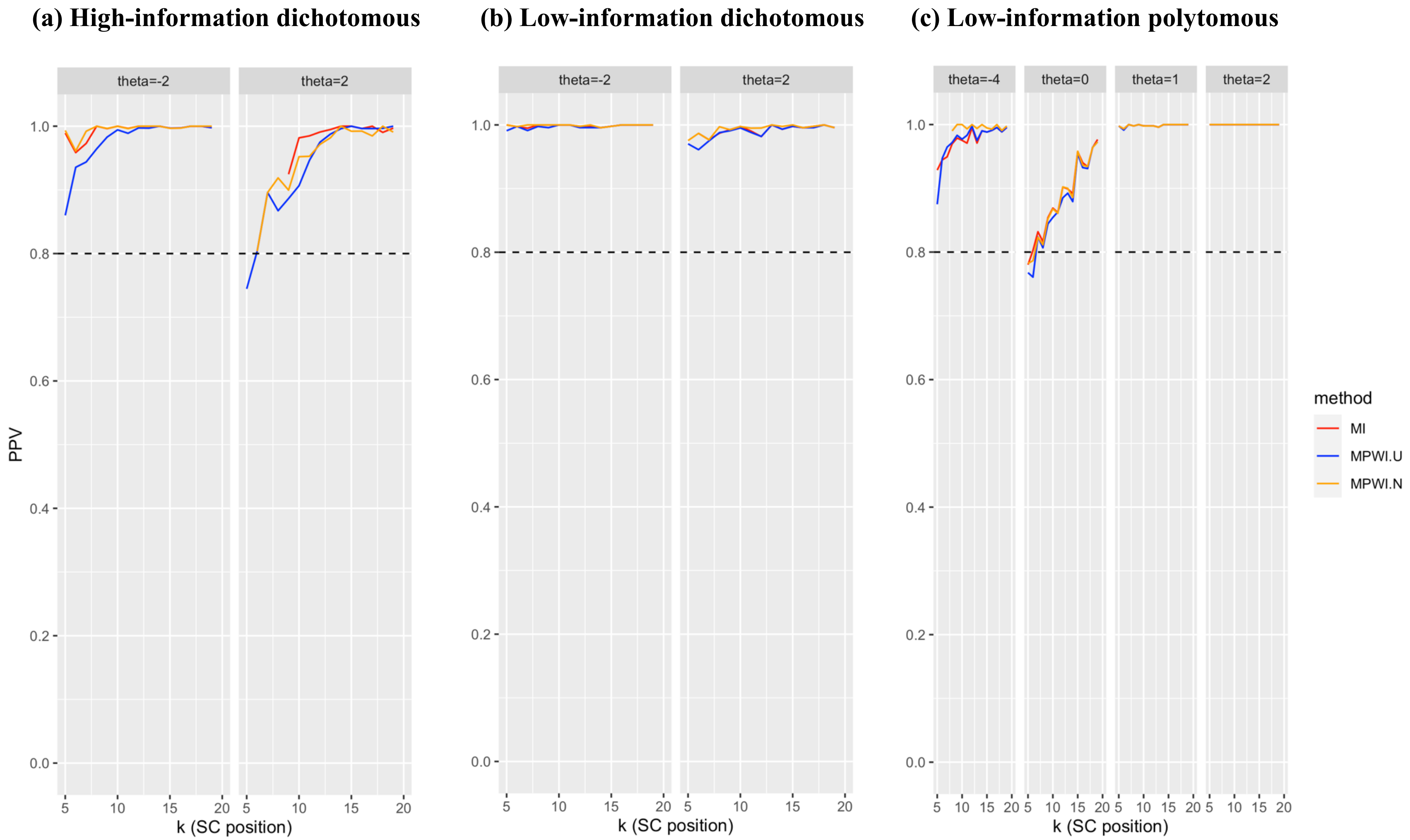
(b) Low-information dichotomous, 20 Items

Method	θ	SC point														
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
	-2	.001	-.026	.047	.041	.014	.036	.029	.044	.044	.026	.035	.045	.017	.009	.005
	-1	.015	.011	.033	.007	.009	.001	.005	.010	.001	.004	.002	.001	.001	.000	.000
MI	0	-.001	-.001	-.003	-.003	.000	-.003	-.004	-.002	-.005	-.002	-.001	-.001	-.001	-.001	-.001
	1	-.001	.003	.004	.005	.003	.003	.005	.001	.000	.001	-.001	.000	.000	-.001	.000
	2	-.079	-.069	-.050	-.032	-.024	-.020	-.012	-.005	.003	.000	.006	.005	.004	.002	.000
	-2	-.125	-.119	-.108	-.089	-.098	-.073	-.060	-.057	-.050	-.044	-.032	-.022	-.017	-.017	-.009
	-1	-.008	-.002	-.001	-.007	-.003	-.005	-.003	-.006	-.005	.001	.000	.000	-.001	-.001	.000
MPWI-N	0	.016	.012	.010	.008	.006	.004	.002	.002	-.001	.001	.001	.001	.000	.000	-.001
	1	.016	.015	.012	.010	.007	.005	.006	.003	.001	.002	.000	.001	.000	-.001	.000
	2	-.061	-.055	-.047	-.043	-.038	-.036	-.030	-.023	-.018	-.017	-.011	-.008	-.006	-.005	-.003
	-2	-.015	-.016	.017	.025	.001	.024	.022	.031	.027	.021	.025	.035	.016	.009	.005
	-1	.044	.043	.043	.028	.024	.016	.017	.015	.008	.011	.008	.004	.003	.002	.001
MPWI-U	0	.038	.028	.021	.015	.012	.009	.005	.005	.001	.002	.002	.001	.000	.000	-.001
	1	.041	.033	.026	.020	.015	.012	.012	.007	.004	.004	.002	.002	.002	.000	.001
	2	.000	-.001	.001	.000	-.002	-.003	.000	.002	.009	.003	.008	.006	.005	.002	.001

(c) Low-information polytomous, 20 Items

Method	θ	SC point														
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
MI	-4	-.011	-.008	-.005	-.004	-.003	-.003	-.002	-.001	-.002	-.001	-.001	-.002	.000	.000	.000
MI	-3	.003	.004	.002	.003	.001	.001	.001	.001	.001	.001	.000	.000	.000	.000	.000
MI	-2	.002	.001	.001	.000	.001	.000	.001	.000	.000	.000	.001	.000	.000	.000	.000
MI	-1	.004	.004	.002	.002	.002	.001	.001	.001	.001	.000	.000	.000	.000	.000	.000
MI	0	.008	.010	.013	.012	.011	.007	.006	.004	.005	.003	.004	.002	.002	.001	.000
MI	1	-.054	-.037	-.034	-.017	-.015	-.009	.001	-.001	.002	.009	.011	.007	.006	.004	.002
MI	2	-.153	-.128	-.099	-.085	-.075	-.061	-.046	-.034	-.017	.001	.007	.016	.021	.026	.023
MPWI-N	-4	-.031	-.028	-.025	-.022	-.020	-.018	-.016	-.013	-.012	-.010	-.009	-.008	-.006	-.004	-.002
MPWI-N	-3	-.005	-.004	-.005	-.004	-.004	-.003	-.003	-.002	-.003	-.002	-.002	-.002	-.001	-.001	-.001
MPWI-N	-2	.005	.004	.003	.002	.001	.001	.001	.001	.000	.001	.001	.000	.000	.000	.000
MPWI-N	-1	.012	.010	.007	.007	.006	.004	.003	.002	.002	.002	.001	.001	.001	.000	.000
MPWI-N	0	.008	.008	.011	.009	.009	.006	.005	.004	.005	.003	.003	.002	.002	.001	.000
MPWI-N	1	-.066	-.055	-.058	-.044	-.040	-.035	-.027	-.024	-.018	-.009	-.006	-.007	-.005	-.002	-.001
MPWI-N	2	-.157	-.140	-.117	-.111	-.104	-.089	-.080	-.066	-.050	-.033	-.023	-.010	.000	.009	.016
MPWI-U	-4	-.018	-.015	-.013	-.012	-.010	-.010	-.008	-.006	-.006	-.005	-.005	-.005	-.003	-.002	-.001
MPWI-U	-3	.002	.003	.002	.002	.001	.001	.001	.001	.001	.001	.000	.000	.001	.000	.000
MPWI-U	-2	.005	.003	.003	.002	.002	.001	.001	.001	.001	.001	.001	.000	.000	.000	.000
MPWI-U	-1	.011	.009	.006	.005	.005	.003	.002	.002	.002	.001	.000	.000	.000	.000	.000
MPWI-U	0	.027	.022	.023	.018	.017	.012	.010	.007	.008	.005	.005	.003	.002	.001	.001
MPWI-U	1	.003	.006	-.004	.005	.004	.004	.009	.007	.008	.013	.013	.009	.007	.005	.003
MPWI-U	2	-.055	-.048	-.039	-.035	-.036	-.024	-.022	-.014	-.003	.011	.014	.021	.025	.027	.024

Figure 5. PPVs of the SC procedure for 20-item CATs


(a) High-information dichotomous

(b) Low-information dichotomous

(c) Low-information polytomous

Table 8. PPVs of the SC procedure for 20-item CATs

(a) High-information dichotomous, 20 Items

Method	θ	SC point														
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
MI	2	.989	.959	.973	1.000	.996	1.000	.997	1.000	1.000	1.000	.997	.997	1.000	1.000	.998
MI	2	N/A	N/A	N/A	N/A	.925	.982	.985	.991	.995	1.000	1.000	.996	1.000	.990	.997
MPWI-N	-2	.993	.961	.992	1.000	.996	1.000	.996	1.000	1.000	1.000	.997	.997	1.000	1.000	1.000
MPWI-	2	N/A	.804	.896	.919	.900	.952	.953	.971	.982	1.000	.992	.992	.985	1.000	.991
MPWI-	-2	.860	.935	.944	.964	.983	.994	.989	.997	.997	1.000	.997	.997	1.000	1.000	.998
MPWI-	2	.745	.804	.896	.867	.886	.907	.947	.975	.988	.996	1.000	.996	.996	.997	1.000

(b) Low-information dichotomous, 20 Items

	θ	SC point														
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
MI	-2	1.000	.997	.995	1.000	1.000	1.000	1.000	.998	.998	.996	.998	1.000	1.000	1.000	1.000
MI	2	N/A	.987	.977	.987	.992	.998	.990	.982	1.000	.993	.998	.996	.998	1.000	.996
MPWI-	-2	1.000	.997	1.000	1.000	1.000	1.000	1.000	.998	1.000	.996	.998	1.000	1.000	1.000	1.000
MPWI-	2	.975	.987	.977	.997	.992	.997	.995	.995	1.000	.998	1.000	.996	.998	1.000	.996
MPWI-	-2	.991	.998	.991	.998	.996	1.000	1.000	.996	.996	.996	.998	1.000	1.000	1.000	1.000
MPWI-	2	.970	.961	.975	.988	.991	.995	.988	.982	1.000	.993	.998	.996	.996	1.000	.996

(c) Low-information polytomous, 20 Items

Method	θ	SC point														
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
MI	-4	.929	.944	.949	.970	.979	.975	.971	.995	.971	.990	.988	.991	.995	.989	.996
MI	0	.780	.803	.832	.817	.854	.869	.862	.902	.899	.892	.958	.940	.934	.964	.977
MI	1	.998	.993	1.000	.998	1.000	.998	.998	.998	.996	1.000	1.000	1.000	1.000	1.000	1.000
MI	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
MPWI-	-4	N/A	N/A	N/A	.990	1.000	1.000	.993	1.000	.994	1.000	.995	.993	1.000	.991	.998
MPWI-	0	.782	.787	.823	.811	.852	.868	.861	.902	.900	.886	.958	.936	.934	.964	.973
MPWI-	1	.998	.993	1.000	.998	1.000	.998	.998	.998	.996	1.000	1.000	1.000	1.000	1.000	1.000
MPWI-	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
MPWI-	-4	.875	.947	.965	.972	.983	.977	.983	.997	.975	.990	.988	.991	.995	.989	.996
MPWI-	0	.768	.761	.824	.807	.844	.854	.863	.885	.892	.879	.954	.932	.931	.964	.973
MPWI-	1	.998	.991	1.000	.998	1.000	.998	.998	.998	.996	1.000	1.000	1.000	1.000	1.000	1.000
MPWI-	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note: N/A means no data.

Figure 6. TPRs of the SC procedure for 20-item CATs

(a) High-information dichotomous **(b) Low-information dichotomous** **(c) Low-information polytomous**

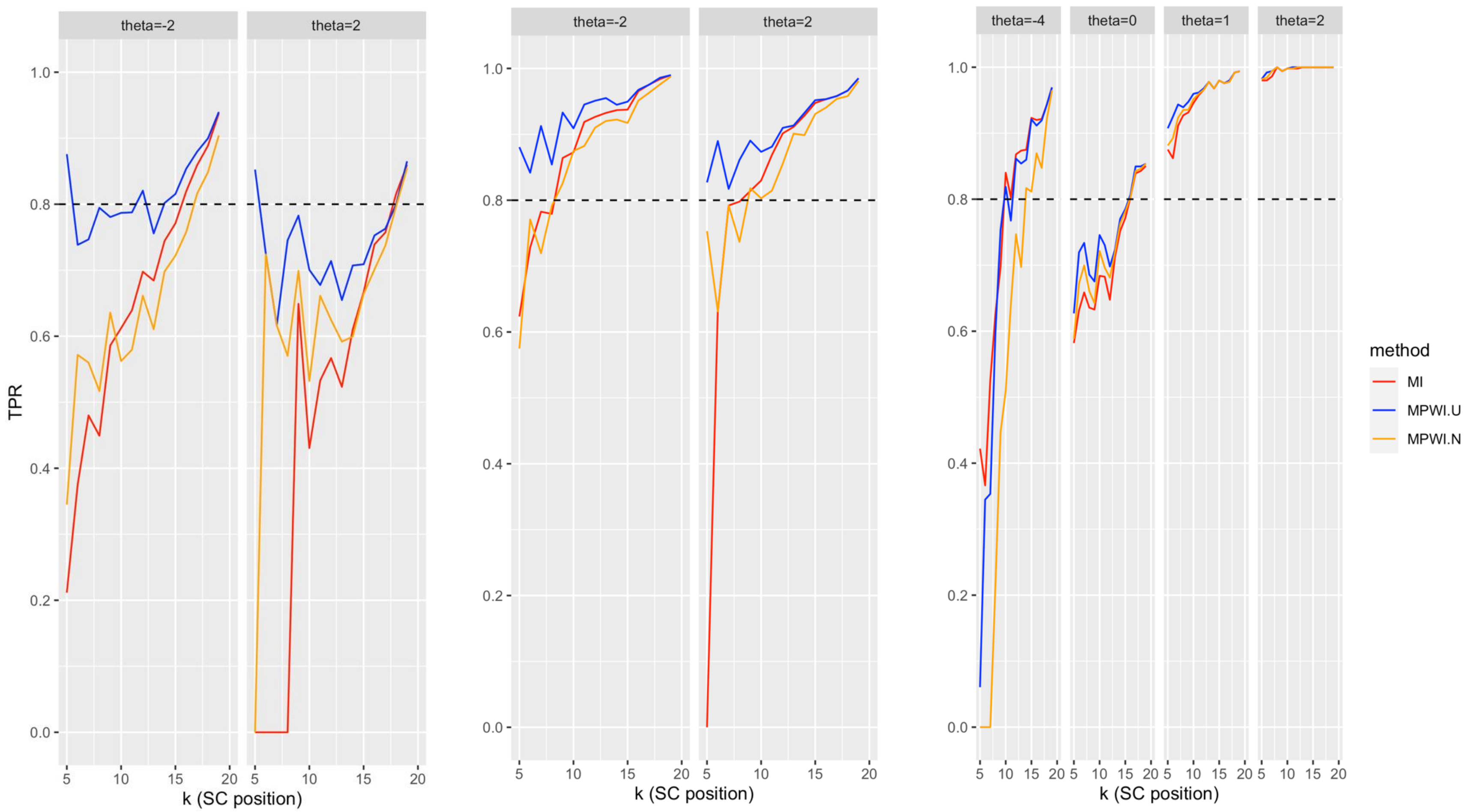


Table 9. TPRs of the SC procedure for 20-item CATs

(a) High-information dichotomous, 20 Items

Method	θ	SC point														
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
MI	-2	.211	.375	.480	.449	.586	.612	.639	.698	.684	.744	.771	.820	.859	.889	.938
MI	2	.000	.000	.000	.000	.649	.430	.533	.567	.523	.610	.667	.739	.757	.816	.860
MPWI-N	-2	.345	.572	.560	.517	.636	.562	.580	.661	.611	.698	.722	.758	.816	.849	.904
MPWI-N	2	.000	.724	.616	.570	.699	.532	.661	.625	.592	.599	.664	.701	.737	.797	.854
MPWI-U	-2	.876	.738	.747	.795	.781	.787	.788	.820	.756	.802	.815	.854	.880	.900	.940
MPWI-U	2	.852	.724	.616	.745	.783	.701	.678	.714	.655	.707	.709	.753	.763	.800	.865

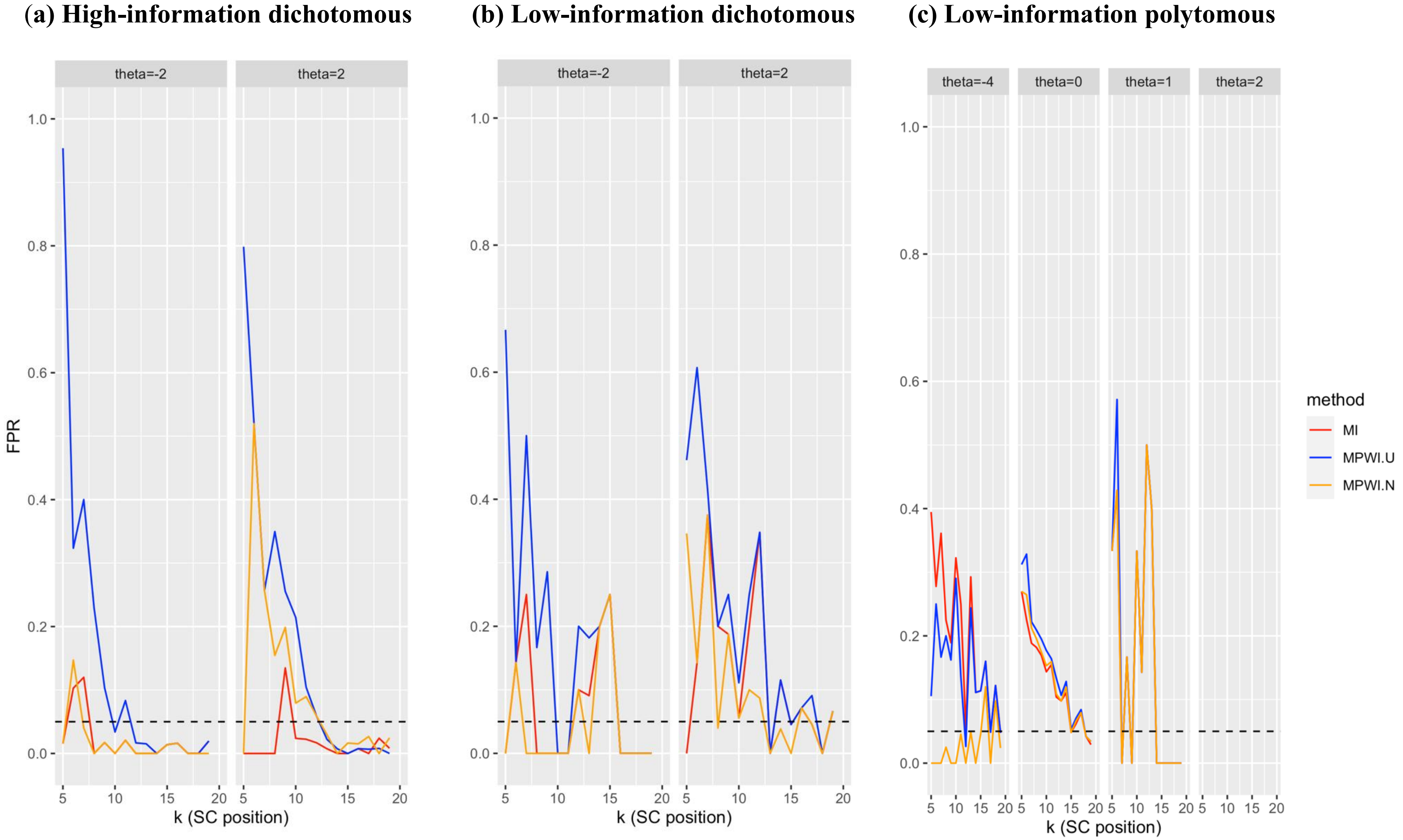
(b) Low-information dichotomous, 20 Items

Method	θ	SC point														
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
MI	-2	.623	.728	.783	.779	.864	.873	.919	.927	.933	.937	.938	.966	.976	.984	.990
MI	2	.000	.631	.792	.798	.814	.830	.869	.901	.911	.928	.948	.953	.958	.966	.985
MPWI-N	-2	.575	.771	.720	.791	.826	.875	.882	.910	.920	.922	.917	.951	.963	.976	.988
MPWI-N	2	.753	.631	.792	.737	.818	.803	.815	.855	.901	.899	.931	.941	.954	.958	.981
MPWI-U	-2	.881	.842	.913	.854	.933	.909	.945	.951	.955	.945	.950	.968	.976	.986	.990
MPWI-U	2	.827	.890	.817	.861	.890	.873	.881	.910	.913	.932	.952	.953	.958	.966	.985

(c) Low-information polytomous, 20 Items

Method	θ	SC point														
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
MI	-4	.422	.366	.526	.626	.698	.840	.803	.868	.874	.875	.923	.920	.922	.943	.969
MI	0	.582	.632	.659	.636	.633	.684	.683	.648	.713	.752	.771	.803	.839	.843	.850
MI	1	.875	.862	.911	.927	.931	.946	.957	.966	.978	.968	.980	.976	.978	.992	.994
MI	2	.980	.980	.986	1.000	.994	.998	.998	.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000
MPWI-N	-4	.000	.000	.000	.215	.447	.510	.640	.747	.697	.817	.811	.869	.847	.922	.965
MPWI-N	0	.588	.672	.700	.661	.643	.722	.696	.681	.720	.763	.782	.803	.843	.846	.854
MPWI-N	1	.881	.892	.923	.935	.935	.952	.959	.966	.978	.968	.980	.976	.978	.992	.994
MPWI-N	2	.982	.984	.994	1.000	.994	.998	.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
MPWI-U	-4	.061	.345	.353	.598	.754	.819	.768	.861	.854	.860	.921	.912	.919	.943	.969
MPWI-U	0	.627	.720	.734	.686	.675	.746	.730	.698	.724	.770	.785	.807	.850	.850	.854
MPWI-U	1	.907	.925	.944	.939	.948	.960	.961	.968	.978	.968	.980	.976	.980	.992	.994
MPWI-U	2	.982	.992	.994	1.000	.994	.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Figure 7. FPRs of the SC procedure for 20-item CATs



(a) High-information dichotomous **(b) Low-information dichotomous** **(c) Low-information polytomous**

Table 10. FPRs of the SC procedure for 20-item CATs

(a) High-information dichotomous, 20 Items

Method	θ	SC point														
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
MI	-2	.015	.103	.120	.000	.017	.000	.021	.000	.000	.000	.014	.016	.000	.000	.020
MI	2	.000	.000	.000	.000	.135	.024	.022	.017	.007	.000	.000	.008	.000	.024	.008
MPWI-N	-2	.015	.147	.040	.000	.017	.000	.021	.000	.000	.000	.014	.016	.000	.000	.000
MPWI-N	2	.000	.520	.257	.154	.199	.079	.090	.059	.030	.000	.016	.015	.027	.000	.025
MPWI-U	-2	.954	.324	.400	.228	.103	.034	.083	.017	.015	.000	.014	.016	.000	.000	.020
MPWI-U	2	.799	.520	.257	.350	.255	.214	.104	.059	.022	.007	.000	.008	.007	.008	.000

(b) Low-information dichotomous, 20 Items

Method	θ	SC point														
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
MI	-2	.000	.143	.250	.000	.000	.000	.000	.100	.091	.200	.250	.000	.000	.000	.000
MI	2	.000	.143	.375	.200	.188	.056	.200	.348	.000	.115	.045	.071	.045	.000	.067
MPWI-N	-2	.000	.143	.000	.000	.000	.000	.000	.100	.000	.200	.250	.000	.000	.000	.000
MPWI-N	2	.346	.143	.375	.040	.188	.056	.100	.087	.000	.038	.000	.071	.045	.000	.067
MPWI-U	-2	.667	.143	.500	.167	.286	.000	.000	.200	.182	.200	.250	.000	.000	.000	.000
MPWI-U	2	.462	.607	.417	.200	.250	.111	.250	.348	.000	.115	.045	.071	.091	.000	.067

(c) Low-information polytomous, 20 Items

Method	θ	SC point														
		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
MI	-4	.395	.278	.361	.225	.189	.323	.250	.053	.293	.111	.114	.160	.049	.122	.048
MI	0	.270	.225	.188	.182	.169	.144	.155	.104	.098	.111	.048	.062	.079	.042	.029
MI	1	.333	.429	.000	.167	.000	.333	.143	.500	.400	.000	.000	.000	.000	.000	.000
MI	2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
MPWI-N	-4	.000	.000	.000	.025	.000	.000	.045	.000	.049	.000	.045	.120	.000	.098	.024
MPWI-N	0	.270	.265	.213	.195	.174	.153	.159	.109	.098	.119	.048	.066	.079	.042	.034
MPWI-N	1	.333	.429	.000	.167	.000	.333	.143	.500	.400	.000	.000	.000	.000	.000	.000
MPWI-N	2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
MPWI-U	-4	.105	.250	.167	.200	.162	.290	.136	.026	.244	.111	.114	.160	.049	.122	.048
MPWI-U	0	.312	.328	.222	.209	.195	.177	.164	.134	.107	.128	.053	.071	.084	.042	.034
MPWI-U	1	.333	.571	.000	.167	.000	.333	.143	.500	.400	.000	.000	.000	.000	.000	.000
MPWI-U	2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Note: N/A means no data.

Reduction in ATL. The results for the reduction in ATL are summarized in Tables 11 and 12. The percentage reduction in ATL was between 54% and 67% under both the high- and low-information dichotomous bank; between 86.6% and 98.0% of the simulees were curtailed for those banks. The percentage reduction in ATL was between 57.7% and 65% under the polytomous bank, and more than 96.6% of the simulees were curtailed. As Tables F1–F3 show, as maximum test length increased, both reduction in ATL and percentage of simulees curtailed increased, though longer maximum test length might artificially boost ATL reduction rates.

Table 11. Reduction in average test length under dichotomous banks

Item bank and test length	$\theta = -2$			$\theta = 2$		
	Reduction in ATL	Reduction in ATL (%)	% simulees SCed	Reduction in ATL	Reduction in ATL (%)	% simulees SCed
HI, 20 items	1.9	54.5	89.0	1.9	54.4	86.6
LI, 20 items	13.5	67.3	97.8	13.4	67.1	98.0

Table 12. Reduction in average test length under the low-information real polytomous bank

Item bank and test length	$\theta = -4$			$\theta = 0$		
	Reduction in ATL	Reduction in ATL (%)	% simulees SCed	Reduction in ATL	Reduction in ATL (%)	% simulees SCed
LI, 20 items	12.8	64.0	99.6	11.5	57.7	96.6

Item bank and test length	$\theta = 1$			$\theta = 2$		
	Reduction in ATL	Reduction in ATL (%)	% simulees SCed	Reduction in ATL	Reduction in ATL (%)	% simulees SCed
LI, 20 items	12.9	64.7	100.0	13.0	65.0	100.0

Discussion and Implementation Guide

The objectives of this study were two-fold. The first objective was to develop and evaluate a procedure to estimate FSEM during a CAT process. Several findings were observed:

1. Under both dichotomous and polytomous item banks, the estimation procedure was satisfactory for the range of true θ within two SDs from its mean. The biases were generally below 0.03 or 0.04, or 10 percent of the termination SEMs, except for when $\theta = -2$ for the dichotomous banks. However, as true θ further deviated from the center, such as when $\theta = 1$ and $\theta = 2$ under the low-information polytomous bank (i.e., 3 and 4 SDs from the mean), the quality of FSEM estimation decreased drastically. This was due to the decreasing bank information as θ became extreme. Therefore, bank information plays a fundamental role in the precision of FSEM estimation.
2. The three item projection methods had comparable performance. MPWI-U performed worse than other methods under the high-information dichotomous bank, but not under the

low-information bank, especially at extreme θ values. Presumably this was because MPWI-U assigned more weight to extreme θ values, which improved estimation precision when the bank information was low. But when bank information was high, this was unnecessary and led to worse results.

3. MTL did not have an obvious impact on estimation precision, meaning that the procedure was robust to maximum test length.

The second objective of the study was to develop and evaluate a procedure to perform SC during a CAT. Results showed that when the true θ was two SDs or above from the center, PPVs, the primary criterion for evaluating the procedure, were mostly above 80% or even close to 100% throughout the tests. This was because (1) PPV was positively related to the base rate of low-precision cases, which were very high at extreme θ values (Table 13); and (2) the SC procedure was highly accurate in making classification decisions at these values. Moreover, PPVs were higher when bank information was lower, such as when comparing the low-information to the high-information dichotomous bank and comparing the more extreme θ values to the less extreme θs under the low-information polytomous bank. The results showed that for simulees with extreme θ values, the low-precision alarms given by the SC procedure are highly trustworthy and warrant the attention of the test administrator. Maximum test length had minor impact on PPVs, meaning that the procedure was robust to maximum test length.

Table 13. Base rates of low-precision cases (%)

θ	High-info dichotomous	Low-info dichotomous	θ	Low-info polytomous
$\theta = -2$	69.2	92.1	$\theta = -4$	91.7
$\theta = 2$	64.7	99.7	$\theta = 0$	61.0
			$\theta = 1$	99.6
			$\theta = 2$	10.0

Secondary criteria, including TPRs and FPRs, were affected by several factors. First, both the TPRs and the FPRs tended to be higher under the low-information bank than the high-information bank. This could be explained by the fact that when information was insufficient, it was more likely that a simulee would be judged as a positive (i.e., a low-precision case), resulting in higher proportions of both true positives and false positives. This can be seen from Tables 11 and 12, which show that the proportions of simulees curtailed (i.e., predicted to be low-precision) were consistently higher under the low-information bank than under the high-information bank. Second, MPWI-U generally had higher TPRs and FPRs than the other two methods at the beginning stages of the test, especially under the two dichotomous banks[1]. This was because MPWI-U weighted item information with the likelihood function, therefore incorporating the uncertainty in θ estima-

[1] An exception was the poor performance of MPWI-U at $\theta = -4$ under the polytomous bank. The main reason for this was that the estimation procedure set the starting θ to be 0. As a result, for the beginning items, the θ estimates were distant from the true θ values. Because the low-information bank peaked around $\theta = -2$, given a negative θ estimate that was distant from $\theta = -4$, it was not surprising that the SC procedure identified the simulee to be a high-precision case, resulting in a false negative. On the contrary, at $\theta = 0$, which equaled the starting θ, the θ estimates were close to 0, resulting in more accurate estimates than at $\theta = -4$.

tion in item projection. In particular, it assigned more weights to extreme θ values when compared to MPWI-N. Because extreme θ values tended to have higher FSEM, the procedure tended to estimate a higher FSEM than the two other methods. This was seen in the MB plots, which showed that MPWI-U tended to have positive MBs, which in turn resulted in higher TPRs and FPRs, as previously discussed. Third, the findings that TPRs were close to 100% at $\theta = 1$ and $\theta = 2$ under the low-information polytomous banks demonstrated a desirable feature of SC, which was that it could achieve perfect classification results at extreme θ values, which was shown by the near 100% curtailment rate (Table 12). The differential performance of the three item projection methods is a topic for further exploration.

To further evaluate the robustness of the SC procedure, supplementary analyses were conducted under more balanced conditions across the θ continuum (−2 to 2) (see Appendix G). These analyses revealed that SC continued to demonstrate meaningful predictive performance even when the base rate of low-precision cases was moderate. For example, at $\theta = -1$ under the high-information bank—where the base rate of low-precision cases was approximately 20%—PPVs increased steadily throughout the test and surpassed 80% by the 10th item. A similar pattern was observed under the low-information bank, where the base rate was approximately 40%: PPVs began near 60% and also exceeded 80% by the 10th item. These trends suggest that the SC procedure's predictive validity was not solely attributable to class imbalance, that is, the predominance of one case type, whether low-precision or high-precision, within the population.

Likewise, TPRs remained moderate to high under these balanced conditions, particularly for MPWI-U, which identified over half of the low-precision cases by mid-test. FPRs, though initially elevated in some cases, declined rapidly to below 5% after 10–12 items. Together, these results support the conclusion that the SC procedure can effectively guide early termination decisions even when low-precision cases are not highly prevalent, and that its utility extends beyond the extreme θ regions emphasized in the main analysis. A more detailed account of these supplementary findings is provided in Appendix G.

The reduction in test length due to SC was high. In general, the procedure was effective in detecting low-precision cases at the early stages of the test and performed curtailment, therefore resulting in substantial reductions in test length—the percentage reduction was mostly in the range of 50 to 70%. Higher base rate of low-precision cases was associated with larger reduction, which was shown in comparing the low-information to the high-information dichotomous bank, as well as comparing the more extreme to the less extreme θ values under the polytomous bank. This finding makes the procedure especially useful when information is low. In addition, the longer the maximum test length, the larger the percentage reduction in test length. A possible reason was that regardless of MTL, it took roughly the same number of starting items to get a reasonably accurate estimate of θ or its likelihood function and thus accurate estimation of FSEM. As a result, the larger the MTL, the more the number of items saved. In addition, larger MTL implied a larger sample size in central limit approximation, making it more accurate.

Overall, the study showed that the central limit approximation yielded good estimation of FSEM. It is the first of its kind in the literature to show that estimation of FSEM during a CAT procedure is possible. Based on the estimation method, an SC procedure was successfully developed to reduce test length with reasonable predictive accuracy.

The present results also contribute to the literature on stochastic curtailment in psychological testing. Finkelman (2008) found that in a sequential mastery testing setting, the proportion of correct decisions (PCD) ranged from 5.4% to 99.3%, depending on how far the true θ deviated from the cutoff θ. PCD was equivalent to PPV in this study. Under the high- and low-information banks, PPVs ranged from approximately 30% to 100% and approximatley 67% to 100% respectively, depending on test length, true θ, item projection method, and SC start point. Note that Finkelman's findings were consistent with this study in that low-information banks had higher PCDs or PPVs than high-information banks. Regarding reduction in test length, Finkelman reported that the stochastically curtailed TSPRT achieved an average reduction of 15% to 20% of test items compared to the TSPRT in an uncurtailed 50-item test, depending on true θ. In this study, reduction in test length ranged from 34.8% to 81.7%, depending on item bank, test length, and true θ. Finkelman (2010) developed some variations of the stochastically curtailed TSPRT method, which further reduced the PCD and the reduction in test length. However, due to the different testing settings between sequential mastery testing and trait estimation, it is difficult to make direct comparisons between this study and Finkelman's and subsequent studies such as Huebner & Fina (2015) and Sie et al. (2015), all of which applied SC to dichotomous classification testing.

Importantly, the SC rule is not intended to replace existing termination methods but to complement them by offering an additional layer of decision support. For instance, if SC predicts that a test is unlikely to meet the target SEM, test administrators can choose to curtail the test, switch to a less stringent stopping rule, or accept reduced precision—depending on the stakes and constraints of the testing context. This flexibility allows SC to support adaptive test design goals while minimizing unnecessary burden on examinees, especially those located in low-information regions of the trait continuum. By incorporating long-range projections into termination logic, SC introduces a practical, anticipatory dimension to the CAT termination framework.

A comparison between SC and traditional termination rules under different item bank conditions—particularly high- versus low-information regions—would be valuable for clarifying the optimal use cases of each method. In high-information regions (e.g., near the center of the θ continuum), most termination rules perform well, and SC might offer limited additional benefit, making it unnecessary to trigger the SC procedure. However, in low-information regions (e.g., two standard deviations from the center of the θ scale), SC excels at identifying likely low-precision cases early in the test. When SC forecasts a low-precision outcome, the administrator might choose to forego SEM-based termination in favor of MI or PSER rules, provided some loss in precision is acceptable. Conversely, if SC predicts a high-precision outcome, then continuing under an SEM rule is advisable. In this way, SC can guide the choice of termination strategy dynamically, balancing efficiency and measurement goals in response to real-time projections. A formal, head-to-head comparison between SC-based compound rules and existing termination rules under matched conditions remains an important direction for future research to quantify their relative advantages and trade-offs across diverse testing scenarios.

The present work aligns with a growing literature on probability-based early stopping rules that aim to reduce respondent burden while maintaining decision quality. For example, Smits and Finkelman (2014) proposed a variable-length testing procedure based on ordinal regression, in which sum scores were forecasted using a proportional odds model and item administration was curtailed once prediction uncertainty dropped below a specified entropy threshold. While their method was designed for settings where sum scores are the primary outcome and IRT-based CAT is inapplicable, it shares conceptual similarities with SC by employing real-time prediction and probabilistic early stopping. In contrast, the SC procedure introduced here operates within an IRT framework and focuses on forecasting whether the final SEM will meet a target, enabling precision-based decisions. Together, these studies illustrate the broader potential of SC logic beyond classification tasks and reinforce the utility of predictive termination rules across measurement paradigms.

Limitations

This study has several limitations. First, the central limit approximation procedure, as the name suggests, was an approximation rather than a strict application of the central limit theorem. In a real item bank, no two items will likely have identical parameters, so the computation of the sum of item information does not conform to the conditions of the central limit theorem. Nevertheless, for practical purposes, interest was in the performance of the procedure even though the conditions were not strictly met.

While the CLT provides a useful approximation for estimating the distribution of the final SEM, its validity depends in part on the number of remaining items ($N - k$) being sufficiently large. In practice, this assumption might not hold when few items remain in the test. Under such circumstances, the sampling distribution of the projected SEM might deviate from normality, potentially affecting the accuracy of forecast-based decisions. Although the current method relies on the CLT for computational efficiency and general applicability, more precise approximations—such as the recursive approach described by Huebner and Finkelman (2016)— might be warranted in cases where the number of remaining items is small. Although their research was in the context of sequential classification tests using the stochastically curtailed sequential probability ratio test, it offers a rigorous framework for computing probabilities under finite item constraints, which could be adapted to enhance SC-based projections in future implementations.

Second, the SC procedure gave a binary judgment of whether the FSEM would likely reach the termination SEM. In some practical applications, SC termination might not be seen as the sole termination rule. Instead, it might be regarded as a procedure to inform the selection of termination rules. As explained above, if during a CAT process, the SC procedure determines that it is a low-precision case, then the test administrator might want to switch to a different termination rule, but with the knowledge that a prespecified value of the SEM might not be reachable by an examinee. Under these circumstances, Wang et al. (2019) recommend a small change in θ estimates repeated across two successive items, as a secondary termination criterion.

Third, the selection of termination SEMs in this study lacked a mathematical basis. The value was determined by observing the histogram of the FSEMs such that a reasonable portion of the simulees would be low-precision cases and the remaining simulees would be high-precision. A different choice of termination SEM will alter the base rates. However, the impact of base rate on results has already been demonstrated and discussed in the current study. Further studies that are designed to replicate and extend this study can experiment with different termination SEMs and different methods to determine appropriate termination SEMs.

Fourth, while this study focused on positive predictive value (PPV), true positive rate (TPR), and false positive rate (FPR), negative predictive value (NPV) was also considered as a potential secondary metric. However, it was decided not to include NPV in the primary analysis because its interpretive value is more limited in the context of test termination decisions. NPV reflects the proportion of examinees *predicted to be high-precision* who actually achieved the target SEM. While informative in some contexts, false negative cases—those misclassified as high-precision when they are actually low-precision—primarily lead to longer tests than necessary. This is a less critical error than false positives, where examinees are incorrectly judged as low-precision and the test is terminated prematurely. Since SC is designed to minimize the likelihood of premature termination, metrics that reflect the cost of false positives (i.e., PPV) are more directly aligned with its goals. However, NPV might be useful in settings where minimizing unnecessary test length is a priority, and its utility warrants further exploration.

Last, but not least, the SC procedure was bank specific. From a practical standpoint, each item bank needs to be calibrated to implement the SC function, which requires time and expertise.

Future Directions

The study can be extended in the following directions: (1) using Bayesian methods instead of MLE as the θ estimation method. Even though Equation 1 holds only for MLE estimators, given the practical nature of this project and the various approximations involved, Bayesian methods might yield better results at least in some respects due to its reduced variance of θ estimates, and (2) the methods can be extended to multidimensional IRT models. In multidimensional CATs, the test administrator might use the desired size of the error ellipse or posterior credibility region as the termination criterion (Reckase, 2009). Methods can be developed to find the distribution of the size of the error ellipse or posterior credibility region, based on which an SC procedure can then be developed.

Implementation Guide

Given the potential of SC in reducing test length, practitioners should consider adding SC features to their CATs. To implement SC with a given item bank, the following steps are needed:

1. Specify a maximum test length. Then run simulations (e.g., 1,000 times at θ levels with low-information) of full-length CATs. Plot histograms of FSEMs at the range of true θ values of interest and determine the appropriate termination SEMs (see Appendix D).
2. Select an item projection method. MPWI-U is recommended with MLE θ estimation due to its favorable performance at extreme θ values, which are usually the region of interest for curtailment.
3. Specify a Type I error rate (e.g., 0.05). Implement the SC procedure once the testing process reaches the SC start point, which is the fifth item in the CAT as discussed in Method above.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika. 43*(4), 561–573. DOI

Ayanlowo, A. O., & Redden, D. T. (2007). Stochastically curtailed phase II clinical trials. *Statistics in Medicine, 26*(7), 1462–1472. DOI

Babcock, B., & Weiss, D. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing*, *1*(1), 1–18. DOI

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29–51. DOI

Bradlow, E. T. (1996). Teacher's Corner: Negative information and the three-parameter logistic model. *Journal of Educational and Behavioral Statistics, 21*(2), 179–185. DOI

Choi, S. W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, *71*(1), 37–53. DOI

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. DOI

Davis, B. R., & Hardy, R. J. (1994). Data monitoring in clinical trials: The case for stochastic curtailment. *Journal of Clinical Epidemiology, 47(9)*, 1033–1042. DOI

DeWeese, J. N., & Weiss, D. J. (2023). *Stochastically curtailed adaptive measurement of change* (2023). Technical report, Computerized Adaptive Testing Lab, University of Minnesota.

Dodd, B. G., Koch, W. R., & Ayala, R. J. D. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13*(2), 129–143. DOI

Dodd, B. G., Koch, W. R., Ayala, R. J. D., & Ayala, R. D. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement, 53*(1), 61–77. DOI

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics, 33*(4), 442–463. DOI

Finkelman, M. (2010). Variations on stochastic curtailment in sequential mastery testing. *Applied Psychological Measurement, 34*(1), 27–45. DOI

Finkelman, M., He, Y., Kim, W., & Lai, A. M. (2011). Stochastic curtailment of health questionnaires: A method to reduce respondent burden. *Statistics in Medicine, 30*(13), 1989–2004. DOI

Finkelman, M., Smits, N., Kim, W., & Riley, B. (2012). Curtailment and stochastic curtailment to shorten the CES-D. *Applied Psychological Measurement, 36*(8), 632–658. DOI

Gialluca, K. A., & Weiss, D. (1979). *Efficiency of an adaptive inter-subset branching strategy in the measurement of classroom achievement*. (Research Report 79-6). University of Minnesota, Department of Psychology, Psychometric Methods Program. WebLink

Halperin, M., Lan, K. K. G., Ware, J. H., Johnson, N. J., & DeMets, D. L. (1982). An aid to data monitoring in long-term clinical trials. *Controlled Clinical Trials, 3,* 311–323. DOI

Huebner, A., & Fina, A. D. (2015). The stochastically curtailed generalized likelihood ratio: A new termination criterion for variable-length computerized classification tests. *Behavior Research Methods, 47*, 549–561. DOI

Huebner, A. R., & Finkelman, M. D. (2016). On computing the key probability in the stochastically curtailed sequential probability ratio test. *Applied Psychological Measurement*, *40*(2), 142–156. DOI

Kim-Kang, G., & Weiss, D. (2008). Adaptive measurement of individual change. *Zeitschrift Fur Psychologie—Journal of Psychology, 216*(1)*,* 49–58. . DOI

Law, M., Grayling, M. J., & Mander, A. (2020). A stochastically curtailed two-arm randomised phase II trial design for binary outcomes. *Pharmaceutical Statistics, 20*(2), 213–228. DOI

Law, M., Grayling, M. J., & Mander, A. (2022). A stochastically curtailed single-arm phase II trial design for binary outcomes. *Journal of Biopharmaceutical Statistics, 3, 2*(5), 671–691. *DOI*

Lee, J. E. (2015). *Hypothesis testing for adaptive measurement of individual change*. Unpublished doctoral dissertation, University of Minnesota.

Lee, J. E. (2019). *Change in K-12 reading achievement on two occasions*. Paper presented at the 2019 Conference of the International Association for Computerized Adaptive Testing, Minneapolis, MN.

Magis, D. (2015). A note on the equivalence between observed and expected information functions with polytomous IRT models. *Journal of Educational and Behavioral Statistics, 40*(1), 96–105. *DOI*

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. *DOI*

Maurelli, V. A., & Weiss, D. (1981). *Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries*. (Research Report 81-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory. *WebLink*

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14(1)*, 59–71. *DOI*

Phadke, C. (2017). *Measuring intra-individual change at two or more occasions with hypothesis testing methods*. Unpublished doctoral dissertation, University of Minnesota.

R Core Team. (2024). *R* [Computer software].

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer New York. *DOI*

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(4), 311–327. *DOI*

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 1, 1–169. DOI*

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika, 38(2), 221–233*. *DOI*

Sie, H., Finkelman, M. D., Bartroff, J., & Thompson, N. A. (2015). Stochastic curtailment in adaptive mastery testing: Improving the efficiency of confidence interval-based stopping rules. *Applied Psychological Measurement*, *39*(4), 278–292. *DOI*

Smits, N., & Finkelman, M. D. (2014). Variable-length testing using the ordinal regression model. *Statistics in Medicine, 33*(3), 488–499. *DOI*

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. Springer New York. *DOI*

Wang, C., Weiss, D. J., & Shang, Z. (2019). Variable-length stopping rules for multidimensional computerized adaptive testing. *Psychometrika*, *84*(3), 749–771. *DOI*

Wang, C., Weiss, D. J., & Suen, K. Y. (2021). Hypothesis testing methods for multivariate multi-occasion intra-individual change. *Multivariate Behavioral Research*, *56*(3), 459–475. *DOI*

Wang, C., Weiss, D. J., Suen, K. Y., Basford, J., & Cheville, A. (2022). Multidimensional computerized adaptive testing: A potential path toward the efficient and precise assessment of applied cognition, daily activity, and mobility for hospitalized patients. *Archives of Physical Medicine and Rehabilitation, 103*(5, Supplement), S3–S14. *DOI*

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25*(4), 317–331. DOI

Ware, J. E., Bjorner, J. B., & Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing: A brief summary of ongoing studies of widely used headache impact scales. *Medical Care, 38*(9), II-73–II-82. DOI

Ware, J. E., Gandek, B., Sinclair, S. J., & Bjorner, J. B. (2005). Item response theory and computerized adaptive testing: Implications for outcomes measurement in rehabilitation. *Rehabilitation Psychology, 50*(1), 71–78. DOI

Ware, J. E., Kosinski, M., Bjornerl, J. B., Bayliss, M. S., Batenhorst, A. S., Tepper, S. J., & Dowson, A. J. (2003). Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research, 12*, 935–952. DOI

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375. DOI

Weiss, D. J. & Şahin, A. (2024). *Computerized adaptive testing: From concept to implementation.* Guilford Press.

Winters, B. D., Bharmal, A., Wilson, R. F., Zhang, A., Engineer, L., Defoe, D., Bass, E. B., Dy, S., & Pronovost, P. J. (2016). Validity of the Agency for Health Care Research and Quality Patient Safety Indicators and the Centers for Medicare and Medicaid Hospital-acquired Conditions: A systematic review and meta-analysis. *Medical Care*, *54*(12), 1105. DOI

Yen, W. M., Burket, G. R., & Sykes, R. C. (1991). Nonunique solutions to the likelihood equation for the three-parameter logistic model. *Psychometrika, 56*(1), 39–54. DOI

Acknowledgments

This publication was adapted from the doctoral thesis of Ming Him Tai. The authors are grateful to Dr. Chun Wang for her technical expertise, which helped resolve numerous mathematical challenges. Funding from the Mayo Clinic supported a pilot study for the dissertation. The authors also acknowledge the David Campbell Fellowship awarded to Ming Him Tai by the University of Minnesota Department of Psychology, which funded his participation in the 2022 International Association for Computerized Adaptive Testing Conference in Frankfurt, Germany, where he presented the pilot study results.

Author Address

mqt5711@psu.edu

Citation

Tai. M. H. DeWeese, J.,N. & Weiss, D .J. (2025). Stochastic curtailment: A new approach to improve efficiency of variable-length computerized adaptive tests.
Journal of Computerized Adaptive Testing, 12(4), 165-235.
DOI: https://10.7333/2508-1204165

Appendix A

The Mean and Variance of Item Information

Consider a dichotomous item under the 3-parameter logistic model (3PL). Denote a_i, b_i, c_i as the discrimination, difficulty and pseudo-guessing parameters, respectively; Denote p_i as the probability of answering the item correctly (or in the keyed direction). Denote u_i as the response to the item, so that $u_i = 1$ when answering correctly, the probability of which is p_i, and $u_i = 0$ when answering incorrectly, the probability of which is $1 - p_i$. The observed information of the item is given by

$$J_i(\theta|u_i) = -\frac{\partial}{\partial\theta^2} lnL_i(\theta|u_i) \quad \text{(A1)}$$

where $L_i(\theta)$ is the likelihood function for the item defined as

$$L_i(\theta|u_i) = p_i^{u_i} \cdot (1 - p_i)^{1-u_i} \quad \text{(A2)}$$

and

$$p_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta-b_i)}}{1 + e^{a_i(\theta-b_i)}} \quad \text{(A3)}$$

Insert Equations A2 and A3 into Equation A1 to obtain the expression for $J_i(\theta|u_i)$ (Bradlow, 1996):

$$J_i(\theta|u_i) = \frac{e^{t_i}}{(1 + e^{t_i})^2} - u_i c_i \frac{e^{t_i}}{(c_i + e^{t_i})^2} \quad \text{(A4)}$$

where $t_i = a_i(\theta - b_i)$.

Dichotomous Response Items

Note that when $c_i = 0, J_i(\theta|u_i) = \frac{e^{t_i}}{(1+e^{t_i})^2}$, meaning that it is a constant with regard to u_i. Therefore, the observed information of a 2-parameter logistic (2PL) item is always equal to its expected information, which is $\frac{e^{t_i}}{(1+e^{t_i})^2}$. This is a well-established conclusion in the literature (Bradlow, 1996; Samejima, 1973; Yen et al., 1991). An important implication is that the SEM estimation method applied in this research applies to the 3PL only, not the 2PL or one-parameter logistic model. Other estimation methods will be needed for these latter models.

In 3PL when $c \neq 0, J_j(\theta|u_i)$ is a function of u_i. Taking the expectation with regard to u_i gives Fisher's expected information (Bradlow, 1996):

$$E_{u_i}[J_i(\theta|u_i)] = E_{u_i}\left[\frac{e^{t_i}}{(1+e^{t_i})^2} - u_i c_i \frac{e^{t_i}}{(c_i+e^{t_i})^2}\right]$$
$$= \frac{e^{t_i}}{(1+e^{t_i})^2} - p_i c_i \frac{e^{t_i}}{(c_i+e^{t_i})^2} \qquad \text{(A5)}$$

Insert the expression of p_i to derive the expression for $\mathrm{I}_i(\hat{\theta}_N|\boldsymbol{u}_k)$,

$$\mathrm{I}_i(\hat{\theta}_N|\boldsymbol{u}_k) = E_{u_i}[J_i(\theta|u_i)] = \left(\frac{e^{t_j}}{1+e^{t_j}}\right)^2 \left(\frac{1-c_j}{c_j+e^{t_j}}\right) \qquad \text{(A6)}$$

The variance of $\mathrm{J_j}(\theta|\mathrm{u_i})$ with regard to $\mathrm{u_i}$ can be calculated as

$$Var_{u_i}[J_i(\theta|u_i)] = Var_{u_i}\left[\frac{e^{t_i}}{(1+e^{t_i})^2} - u_i c_i \frac{e^{t_i}}{(c_i+e^{t_i})^2}\right]$$
$$= Var_{u_i}\left[-u_i c_i \frac{e^{t_i}}{(c_i+e^{t_i})^2}\right]$$
$$= \left[-c_j \frac{e^{t_j}}{\left(c_j+e^{t_j}\right)^2}\right]^2 p_j(1-p_j) \qquad \text{(A7)}$$

Insert the expression of p_i to derive:

$$Var_{u_i}[J_i(\theta|u_i)] = \frac{c_j^2 e^{2t_j}(c+e^{t_j})(1-c)}{\left(c_j+e^{t_j}\right)^4 (1+e^{t_j})^2} \qquad \text{(A8)}$$

Polytomous Response Items

There are two broad classes of polytomous IRT models: difference models and divide-by-total models (Magis, 2015). Difference models encompass the graded response model (GRM; Samejima, 1968) and the modified graded response model (Muraki, 1990). Divide-by-total models encompasses the partial credit model (PCM; Masters, 1982), the generalized PCM (Muraki, 1990), the rating scale model (Andrich, 1978a, 1978b), and the nominal response model (Bock, 1972), among others. Importantly, observed information and Fisher's expected information are completely equivalent under divide-by-total models, but different under difference models (Magis, 2015). Therefore, the CLT-based estimation method works for the difference models only. The present research uses the GRM because it is the most commonly used difference model.

Consider a polytomous item i. Let $g_i + 1$ be the number of response categories. Denote X_i as the item score, where $X_i \in \{0,1,\ldots,g_i\}$. Denote $P_{ik}(\theta) = \Pr\,(X_i = \mathrm{k}|\theta)$ as the probability of score k $(k = 0,1,\ldots,g_i)$ for an examinee with ability level θ. Under the GRM, $P_{ik}(\theta) = P^*_{ik}(\theta) - P^*_{i(k+1)}(\theta)$, where $P^*_{ik}(\theta)$ is the cumulative probability of the examinee scoring k or a score above,

$P_{ik}^{*}(\theta) = Pr\{X_i \geq \text{k}|\theta\}$. Note that $P_{ik}^{*}(\theta)$ can be expressed using the 2PL model in the dichotomous case, i.e.,

$$P_{ik}^{*}(\theta) = \frac{e^{a_i(\theta - b_{ik})}}{1 + e^{a_i(\theta - b_{ik})}} \tag{A9}$$

where a_i is the discrimination parameter for item, b_{ik} is the threshold (difficulty) parameter for category k. The observed information of the item can be calculated using the same procedure as the dichotomous case, which gives the following result (Magis, 2015):

$$J_i(\theta) = \sum_{k=0}^{g_i} \tau_{ik} \left[\frac{p_{ik}'(\theta)^2}{p_{ik}(\theta)^2} - \frac{p_{ik}''(\theta)}{p_{ik}(\theta)} \right] \tag{A10}$$

where τ_{ik} is the indicator being equal to 1 if $X_i = \text{k}$ and 0 otherwise. $P_{ik}'(\theta)$ and $P_{ik}''(\theta)$ are the first and the second derivatives of $P_{ik}(\theta)$, respectively. Take the expectation with respect to τ_{ik} to derive Fisher's expected information (Magis, 2015):

$$\begin{aligned} \text{I}_i\left(\hat{\theta}_N|\boldsymbol{u}_k\right) = E_{\tau_{ik}}[J_i(\theta)] &= E_{\tau_{ik}}\left[\sum_{k=0}^{g_i} \tau_{ik} \left[\frac{p_{ik}'(\theta)^2}{p_{ik}(\theta)^2} - \frac{p_{ik}''(\theta)}{p_{ik}(\theta)} \right] \right] \\ &= \sum_{k=0}^{g_i} E[\tau_{ik}] \left[\frac{p_{ik}'(\theta)^2}{p_{ik}(\theta)^2} - \frac{p_{ik}''(\theta)}{p_{ik}(\theta)} \right] \\ &= \sum_{k=0}^{g_i} P_{ik}(\theta) \left[\frac{p_{ik}'(\theta)^2}{p_{ik}(\theta)^2} - \frac{p_{ik}''(\theta)}{p_{ik}(\theta)} \right] \\ &= \sum_{k=0}^{g_i} \left[\frac{p_{ik}'(\theta)^2}{p_{ik}(\theta)} - p_{ik}''(\theta) \right]. \end{aligned} \tag{A11}$$

It is easier to derive the variance of $J_i(\theta)$ by applying the formula $Var\left(J_i(\theta)\right) = E[J_i^2(\theta)] - \{E[J_i(\theta)]\}^2$, where $E[J_i(\theta)]$ has been derived in the equation above. To derive $E[J_i^2(\theta)]$:

$$J_i^2(\theta) = \left(\sum_{k=0}^{g_i} \tau_{ik} \left[\frac{p_{ik}'(\theta)^2}{p_{ik}(\theta)^2} - \frac{p_{ik}''(\theta)}{p_{ik}(\theta)} \right] \right)^2$$

$$\begin{aligned}
&= \sum_{k=0}^{g_i} \tau_{ik}^2 \left(\frac{p'_{ik}(\theta)^2}{p_{ik}(\theta)^2} - \frac{p''_{ik}(\theta)}{p_{ik}(\theta)}\right)^2 \\
&+ \sum_{m\neq n}^{g_i} 2\tau_{im}\tau_{in} \left(\frac{p'_{im}(\theta)^2}{p_{im}(\theta)^2} - \frac{p''_{im}(\theta)}{p_{im}(\theta)}\right)\left(\frac{p'_{in}(\theta)^2}{p_{in}(\theta)^2} - \frac{p''_{in}(\theta)}{p_{in}(\theta)}\right)
\end{aligned} \tag{A12}$$

Note that $\tau_{jm}\tau_{jn} = 0$ for $m \neq n$, because only one of the g_j response categories will be endorsed by the examinee, i.e., takes the value of 1; All other response categories will take the value of 0. So $\tau_{jm}\tau_{jn} = 1 \times 0$ or $0 \times 0 =$. Therefore, all items involving $\tau_{jm}\tau_{jn}$ are zero. In addition, note that $\tau_{jk}^2 = \tau_{jk}$ because τ_{jk} is a binary variable. Therefore,

$$\begin{aligned}
E[J_i^2(\theta)] &= E\left[\sum_{k=0}^{g_i} \tau_{ik}^2 \left(\frac{p'_{ik}(\theta)^2}{p_{ik}(\theta)^2} - \frac{p''_{ik}(\theta)}{p_{ik}(\theta)}\right)^2\right] \\
&= E\left[\sum_{k=0}^{g_i} \tau_{ik} \left(\frac{p'_{ik}(\theta)^2}{p_{ik}(\theta)^2} - \frac{p''_{ik}(\theta)}{p_{ik}(\theta)}\right)^2\right] \\
&= \sum_{k=0}^{i} E[\tau_{ik}] \left(\frac{p'_{ik}(\theta)^2}{p_{ik}(\theta)^2} - \frac{p''_{ik}(\theta)}{p_{ik}(\theta)}\right)^2 \\
&= \sum_{k=0}^{g_i} p_{ik}(\theta) \left(\frac{p'_{ik}(\theta)^2}{p_{ik}(\theta)^2} - \frac{p''_{ik}(\theta)}{p_{ik}(\theta)}\right)^2
\end{aligned} \tag{A13}$$

and so,

$$\begin{aligned}
&Var\big(J_i(\theta)\big) = E[J_i^2(\theta)] - \{E[J_i(\theta)]\}^2 \\
&= \sum_{k=0}^{g_j} p_{jk}(\theta) \left(\frac{p'_{jk}(\theta)^2}{p_{jk}(\theta)^2} - \frac{p''_{jk}(\theta)}{p_{jk}(\theta)}\right)^2 - \left(\sum_{k=0}^{g_i}\left[\frac{p'_{ik}(\theta)^2}{p_{ik}(\theta)} - p''_{ik}(\theta)\right]\right)^2
\end{aligned} \tag{A14}$$

Appendix B

The Empirical Mean and Variance of the Observed Information of a Test

The empirical mean of the observed information can be expressed as

$$\begin{aligned} E\left[J\left(\hat{\theta}_N|\boldsymbol{u}_k\right)\right] &= E\left[\sum_{i=1}^{k} J_i\left(\hat{\theta}_N|\boldsymbol{u}_k\right) + \sum_{i=k+1}^{N} J_i\left(\hat{\theta}_N|\boldsymbol{u}_k\right)\right] \\ &= \sum_{i=1}^{k} J_i\left(\hat{\theta}_N|\boldsymbol{u}_k\right) + \sum_{i=k+1}^{N} I_i\left(\hat{\theta}_N|\boldsymbol{u}_k\right), \end{aligned} \tag{B1}$$

where $J_i\left(\hat{\theta}_N|\boldsymbol{u}_k\right), i = 1,2,\dots,k$ denotes the information of an administered item evaluated at $\hat{\theta}_N$, and $I_i\left(\hat{\theta}_N|\boldsymbol{u}_k\right) = E_{u_i}\left[J_i\left(\hat{\theta}_N|\boldsymbol{u}_k\right)\right], i = k+1, k+2, \dots, N$ denotes Fisher's expected information of an item yet to be administered evaluated at $\hat{\theta}_N$. Note that Fisher's expected information of an item is the expected value of the observed information of the item, treating the item response u_i as a random variable. Since the first k items have been administered, their response pattern is known, so their observed information is a known constant, and the expectation of a constant is the constant itself. The remaining $N-k$ items are not administered yet, so their response patterns as well as their item parameters are unknown, and their observed information is also unknown. Taking the expectation of the observed information of these items with respect to their responses gives their Fisher expected information (van der Linden & Glas, 2010). Equation 3 says that the empirical mean of the observed information of the test equals the total observed information of the first k items and the total Fisher's expected information of the remaining $N-k$ items.

The empirical variance of the observed information of the test can be expressed as

$$\begin{aligned} Var\left[J\left(\hat{\theta}_N|\boldsymbol{u}_k\right)\right] &= Var\left[\sum_{i=1}^{k} J_i\left(\hat{\theta}_N|\boldsymbol{u}_k\right) + \sum_{i=k+1}^{N} J_i\left(\hat{\theta}_N|\boldsymbol{u}_k\right)\right] \\ &= Var\left[\sum_{i=k+1}^{N} J_i\left(\hat{\theta}_N|\boldsymbol{u}_k\right)\right] \left(\text{because } \sum_{i=1}^{k} J_i\left(\hat{\theta}_N|\boldsymbol{u}_k\right) \text{ is a constant}\right) \\ &= \sum_{i=k+1}^{N} Var\left[J_i\left(\hat{\theta}_N|\boldsymbol{u}_k\right)\right] \text{ (assuming local independence)} \\ &= \sum_{i=k+1}^{N} \left\{E\left[J_i^2\left(\hat{\theta}_N|\boldsymbol{u}_k\right)\right] - \left\{E\left[J_i\left(\hat{\theta}_N|\boldsymbol{u}_k\right)\right]\right\}^2\right\} \text{ (by the definition of variance)} \end{aligned}$$

$$= \sum_{i=k+1}^{N} \left\{ E\left[J_i^2\left(\hat{\theta}_N | \boldsymbol{u}_k\right)\right] - \left[I_i\left(\hat{\theta}_N | \boldsymbol{u}_k\right)\right]^2 \right\}. \qquad \text{(B2)}$$

Equation B.2 states that the empirical variance of the information of the test equals the sum of the individual variances of the information of the remaining $N - k$ items. The information of the administered k items has no variance because it is a constant.

There are two sets of unknowns in Equations B1 and B2: $I_i\left(\hat{\theta}_N | \boldsymbol{u}_k\right)$ and $E\left[J_i^2\left(\hat{\theta}_N | \boldsymbol{u}_k\right)\right]$ for $i = k + 1, k + 2, \ldots, N$. There are two obstacles to estimating them: First, after k administered items, the remaining $N - k$ items are unknown, so an item projection method is needed to select $N - k$ items from the item bank; second, $\hat{\theta}_N$ is unknown, so it needs to be estimated. There are three approaches to resolve these issues(see Appendix C).

Appendix C

Item Projection Methods

Approach 1: Maximum information. This is the most straightforward approach. $\hat{\theta}_N$ is estimated using $\hat{\theta}_k$, which is the trait level estimate after k administered items. Then, the $N-k$ most informative items evaluated at $\hat{\theta}_k$ from the unused items in the bank are selected. Their information is computed as follows:

$$I_i(\hat{\theta}_N|\boldsymbol{u}_k) = I_i(\hat{\theta}_k) \tag{C1}$$

$$E[J_i^2(\hat{\theta}_N|\boldsymbol{u}_k)] = E[J_i^2(\hat{\theta}_k)]\,. \tag{C2}$$

where $i \in \{all\ remaining\ items\ in\ the\ item\ bank\}$.

Approach 2: Maximum posterior-weighted information with uniform prior. An apparent limitation of the previous approach is that $\hat{\theta}_k$ is used as a point estimate of θ, which does not consider the uncertainty in estimation. This can be improved by using posterior-weighted information, which obviates the need for a point estimate of θ:

$$I_i(\hat{\theta}_N|\boldsymbol{u}_k) = \int_{-\infty}^{+\infty} g(\theta|\boldsymbol{u}_k)I_i(\theta)\,d\theta \tag{C3}$$

$$E[J_i^2(\hat{\theta}_N|\boldsymbol{u}_k)] = \int_{-\infty}^{+\infty} g(\theta|\boldsymbol{u}_k)E[J_i^2(\theta)]\,d\theta \tag{C4}$$

where $g(\theta|\boldsymbol{u}_k)$ is the posterior distribution of θ after administering k items. By the Bayes theorem,

$$g(\theta|\boldsymbol{u}_k) = \frac{L(\theta|\boldsymbol{u}_k)g(\theta)}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{u}_k)g(\theta)d\theta} \tag{C5}$$

where $L(\theta|\boldsymbol{u}_k)$ is the likelihood function of θ given $\boldsymbol{u}_k$, $g(\theta)$ is the prior distribution of θ. The $N-k$ items with the largest posterior-weighted information will be selected.

Two commonly used prior distributions were considered: the uniform distribution and the normal distribution. The uniform distribution places an equal weight across the regions of θ, whereas the normal distribution places the most weight at the center of the region and a decreasing amount of weight as θ deviates from the center. As a result, relatively speaking, using a uniform distribution is conducive to estimating examinees with extreme θ values, and using a normal distribution is conducive to estimating examinees with true θ values around the center of the distribution.

Approach 2 used a uniform distribution. Assume that $g(\theta) = \frac{1}{p-q}, \theta \in [p, q]$, i.e., θ follows a uniform distribution. Insert $g(\theta)$ into Equation C5:

$$g(\theta|\boldsymbol{u}_k) = \frac{L(\theta|\boldsymbol{u}_k)\dfrac{1}{p-q}}{\int L(\theta|\boldsymbol{u}_k)\dfrac{1}{p-q}d\theta} = \frac{L(\theta|\boldsymbol{u}_k)}{\int_q^p L(\theta|\boldsymbol{u}_k)d\theta}. \tag{C6}$$

I

Insert Equation C6 into Equation C3:

$$I_i(\hat{\theta}_N|\boldsymbol{u}_k) = \int_q^p \frac{L(\theta|\boldsymbol{u}_k)}{\int_q^p L(\theta|\boldsymbol{u}_k)d\theta} I_i(\theta)\, d\theta = \frac{\int_q^p L(\theta|\boldsymbol{u}_k)I_i(\theta)d\theta}{\int_q^p L(\theta|\boldsymbol{u}_k)d\theta}. \tag{C7}$$

Insert Equation C7 into Equation C4:

$$\begin{aligned} E\left[J_i^2(\hat{\theta}_N|\boldsymbol{u}_k)\right] = \int\limits_{-\infty}^{+\infty} g(\theta|\boldsymbol{u}_k)E[J_i^2(\theta)]\, d\theta = \int_q^p \frac{L(\theta|\boldsymbol{u}_k)}{\int_q^p L(\theta|\boldsymbol{u}_k)d\theta} E[J_i^2(\theta)]d\theta \\ = \frac{\int_q^p L(\theta|\boldsymbol{u}_k)E[J_i^2(\theta)]d\theta}{\int_q^p L(\theta|\boldsymbol{u}_k)d\theta}. \end{aligned} \tag{C8}$$

In the simulations, set $p = -4$ and $q = 4$. This was the range used in the θ estimation algorithm in this study.

Approach 3: Maximum posterior-weighted information with a normal prior. Assume that $g(\theta) = \frac{1}{\sqrt{2\pi}}(-\frac{\theta^2}{2}), \theta \in (-\infty, +\infty)$, i.e., θ follows a standard normal distribution. Insert $g(\theta)$ into Equation C5:

$$g(\theta|\boldsymbol{u}_k) = \frac{L(\theta|\boldsymbol{u}_k)\dfrac{1}{\sqrt{2\pi}}exp(-\dfrac{\theta^2}{2})}{\int L(\theta|\boldsymbol{u}_k)\dfrac{1}{\sqrt{2\pi}}exp(-\dfrac{\theta^2}{2})d\theta} = \frac{L(\theta|\boldsymbol{u}_k)exp(-\dfrac{\theta^2}{2})}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{u}_k)exp(-\dfrac{\theta^2}{2})d\theta}. \tag{C9}$$

Insert Equation C9 into Equation C3:

$$\begin{aligned} I_i(\hat{\theta}_N|\boldsymbol{u}_k) = \int_{-\infty}^{+\infty} \frac{L(\theta|\boldsymbol{u}_k)exp(-\dfrac{\theta^2}{2})}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{u}_k)exp(-\dfrac{\theta^2}{2})d\theta} I_i(\theta)\, d\theta \\ = \frac{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{u}_k)I_i(\theta)exp(-\dfrac{\theta^2}{2})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{u}_k)exp(-\dfrac{\theta^2}{2})d\theta}. \end{aligned} \tag{C10}$$

Insert Equation C10 into Equation C4:

$$
\begin{aligned}
E\left[J_i^2\left(\hat{\theta}_N|\boldsymbol{u}_k\right)\right] &= \int_{-\infty}^{+\infty} g(\theta|\boldsymbol{u}_k)E[J_i^2(\theta)]\,d\theta \\
&= \int_{-\infty}^{+\infty} \frac{L(\theta|\boldsymbol{u}_k)exp(-\frac{\theta^2}{2})}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{u}_k)exp(-\frac{\theta^2}{2})d\theta} E[J_i^2(\theta)]d\theta \\
&= \frac{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{u}_k)E[J_i^2(\theta)]exp(-\frac{\theta^2}{2})d\theta}{\int_{-\infty}^{+\infty} L(\theta|\boldsymbol{u}_k)exp(-\frac{\theta^2}{2})d\theta}.
\end{aligned} \tag{C81}
$$

The calculations of $I_i(\theta)$ and $E[J_i^2(\theta)]$ for a 3PL dichotomous response model and a polytomous response model are presented in Appendix B.

Appendix D

Determining the N and τ for Each Item Bank

The method of determining the N and τ for each item bank was as follows. For each item bank, full-length CATs were replicated 1,000 times at each of the true θ levels specified. Three values of N, which were 20, 30, and 40 were investigated for each dichotomous bank. Three values of N, which were 15, 20, and 25, were investigated for the polytomous bank. Figures E1–E3 display the histograms of the FSEMs for each item bank under each N and each θ level from the simulations. To create scenarios in which low-precision cases occurred due to insufficient test information at extreme θ levels, the appropriate τ for a given test length should be selected so that the majority of the replications at and around the center of the θ continuum were below τ (i.e., high-precision cases), while the majority of the replications at and around the extremes of the θ continuum were above τ (i.e., low-precision cases). Based on this principle, the τs were selected by observing the histograms in Figures E1–E3.

Appendix E

Supplementary Figures

Figure E1. Histograms of FSEMs by maximum test length (by column) and by θ (by row) under the high-information dichotomous item bank

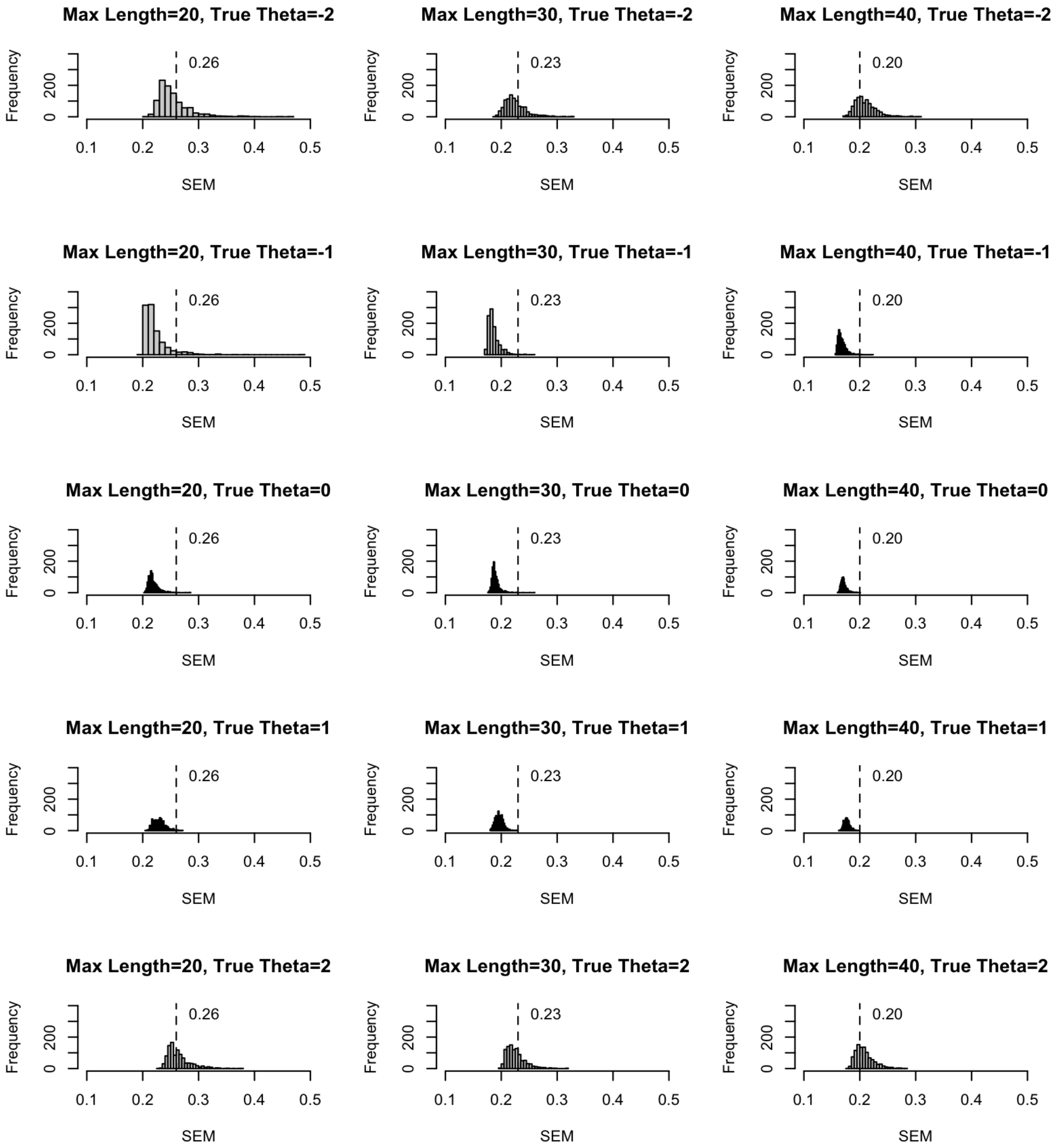


Figure E2. Histograms of FSEMs by maximum test length (by column) and by θ (by row) under the low-information dichotomous item bank

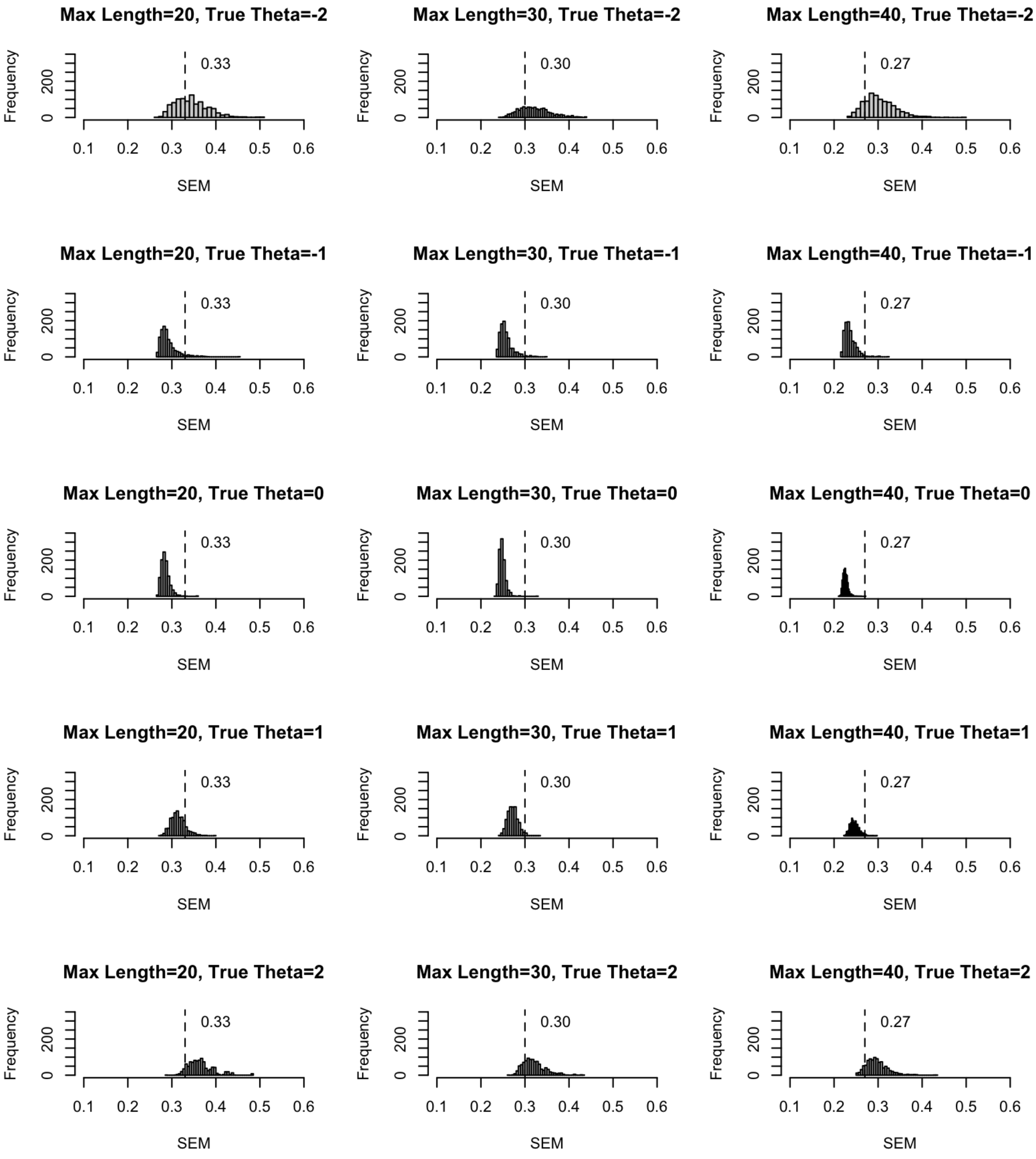


Figure E3. Histograms of FSEMs by maximum test length (by column) and by θ (by row) under the low-information polytomous bank

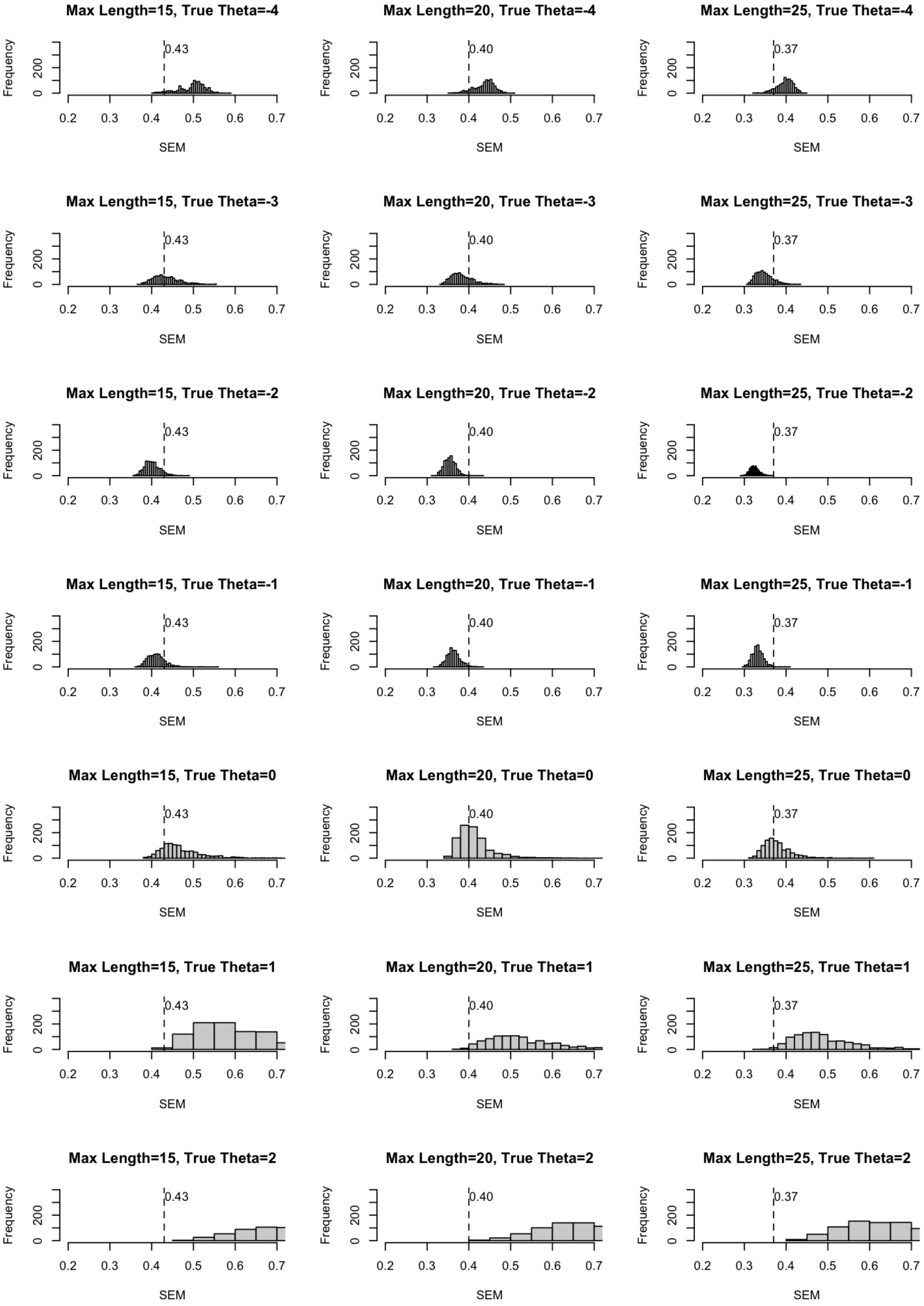
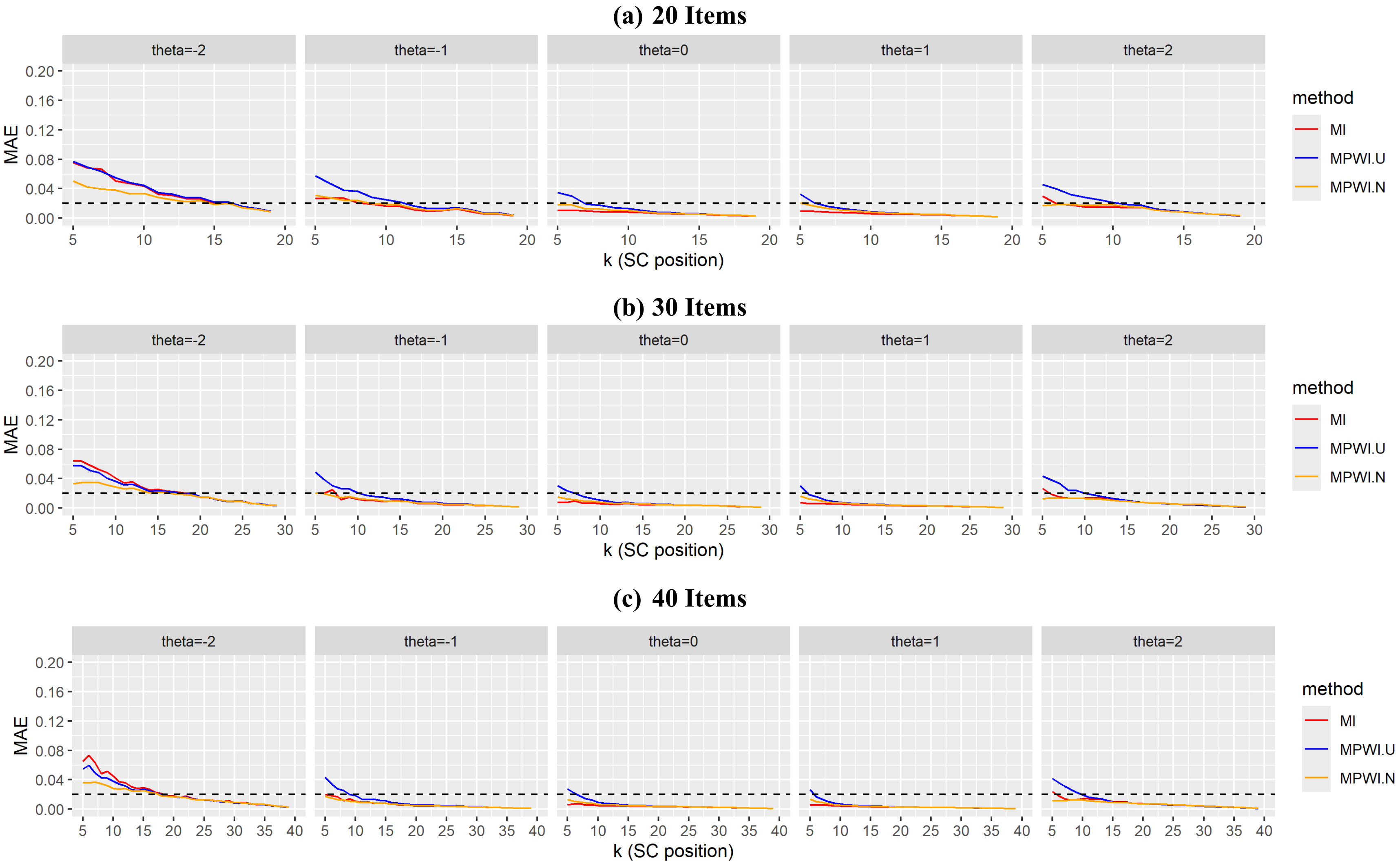


Figure E4. MAE of the FSEM estimation procedure under the high-information dichotomous bank

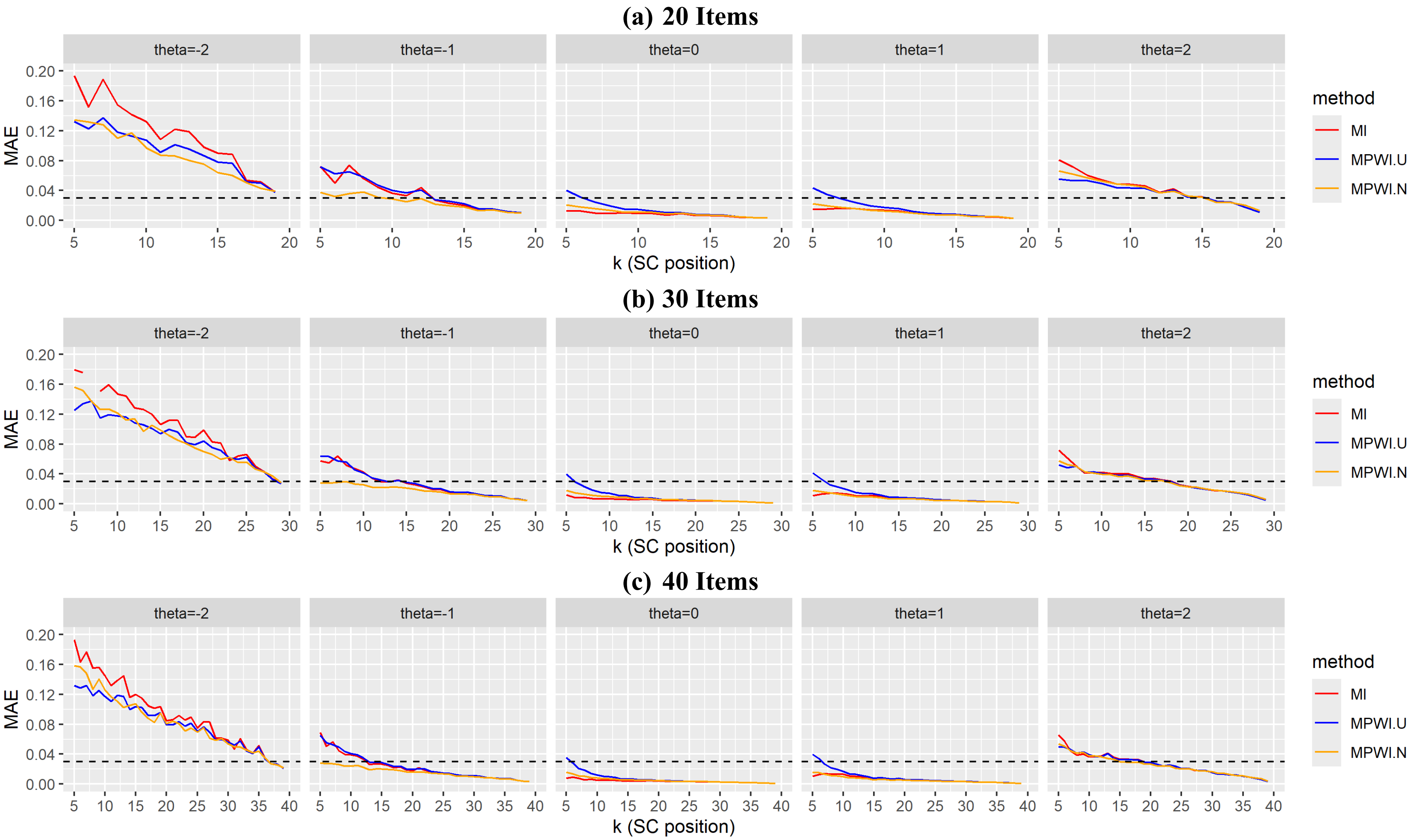


(a) 20 Items

(b) 30 Items

(c) 40 Items

Figure E5. MAE of the FSEM estimation procedure under the low-information dichotomous bank

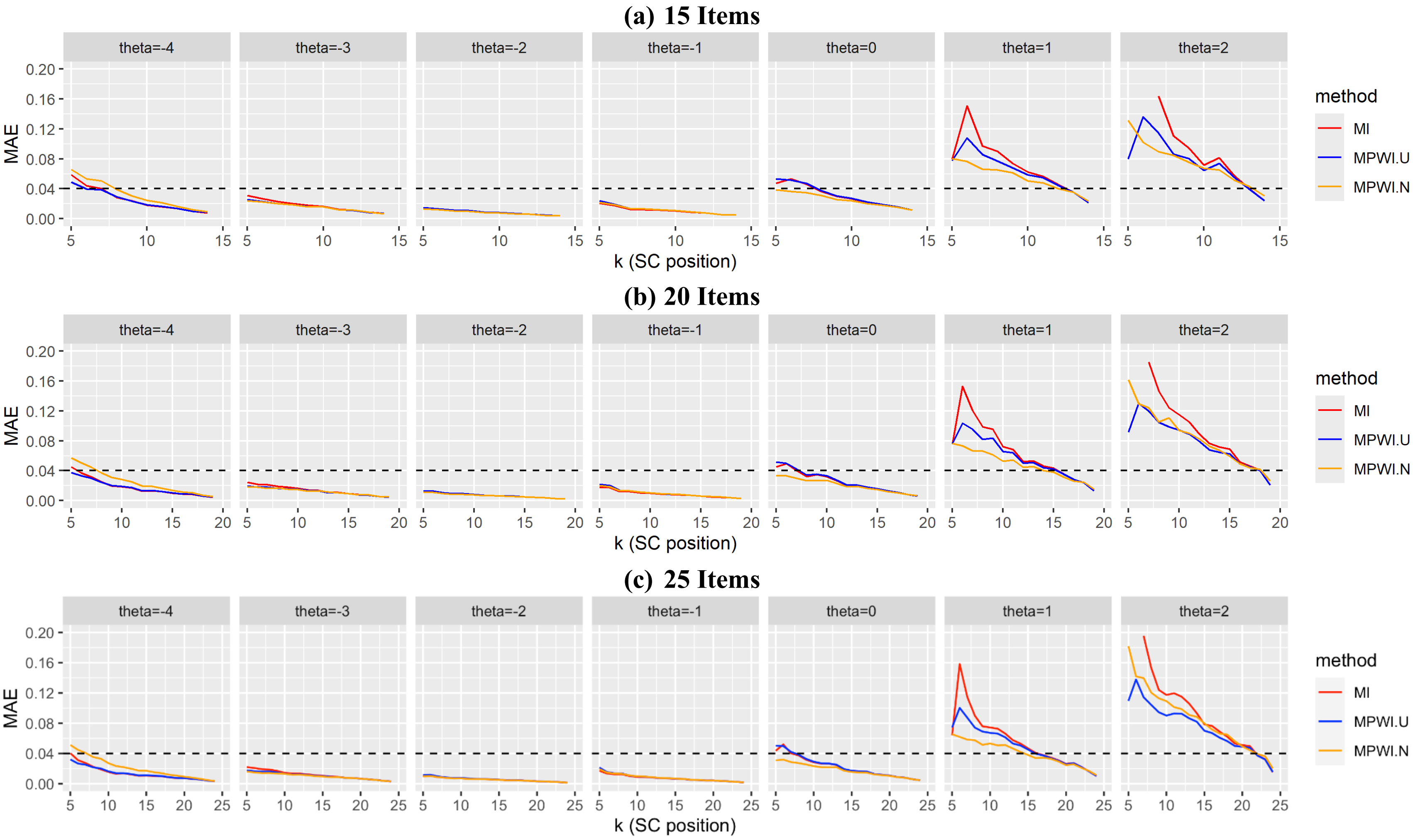


(a) 20 Items

(b) 30 Items

(c) 40 Items

Figure E6. MAE of the FSEM estimation procedure under the low-information polytomous bank

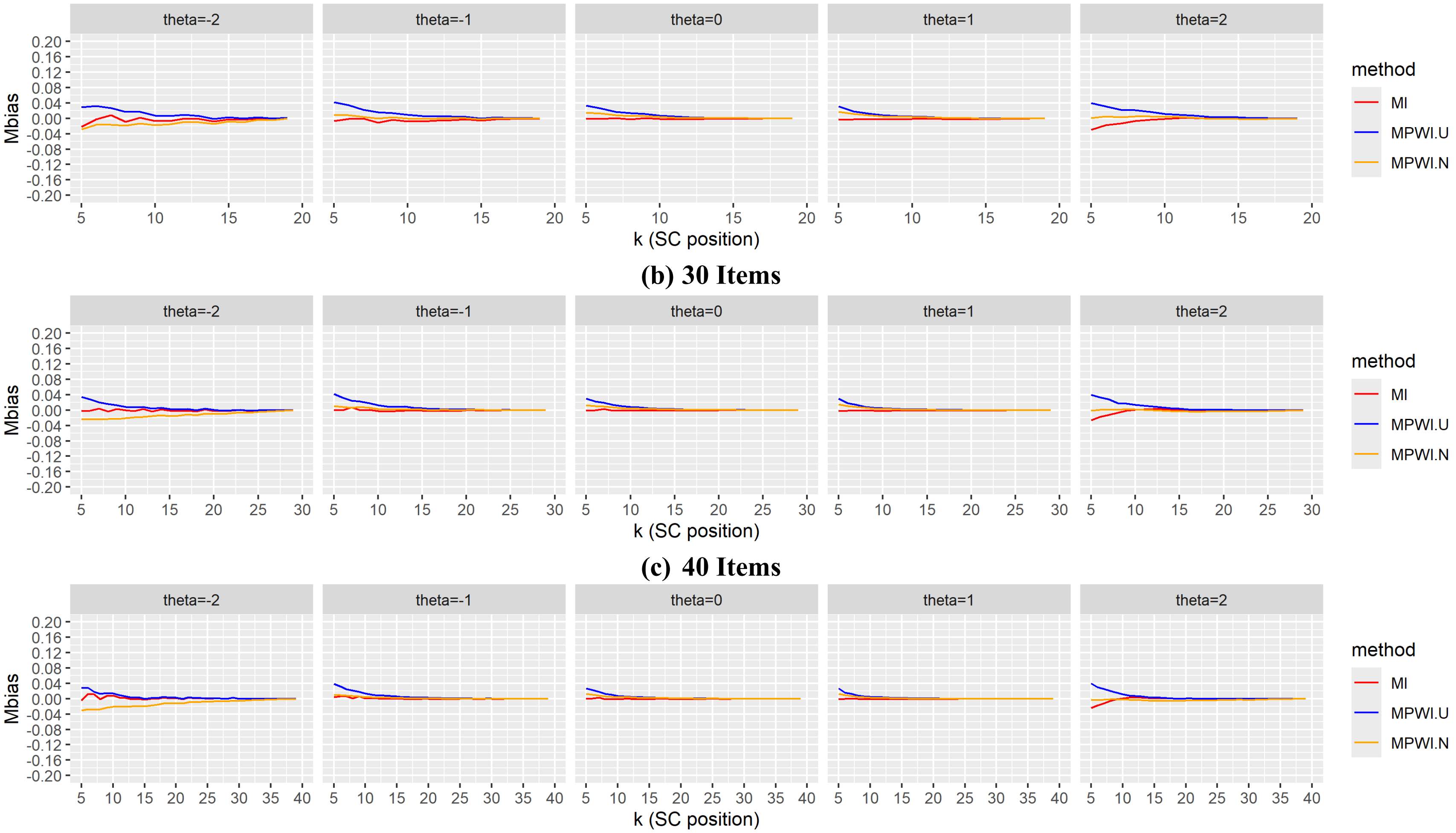


(a) 15 Items

(b) 20 Items

(c) 25 Items

Figure E7. MB of the FSEM estimation procedure under the high-information dichotomous bank

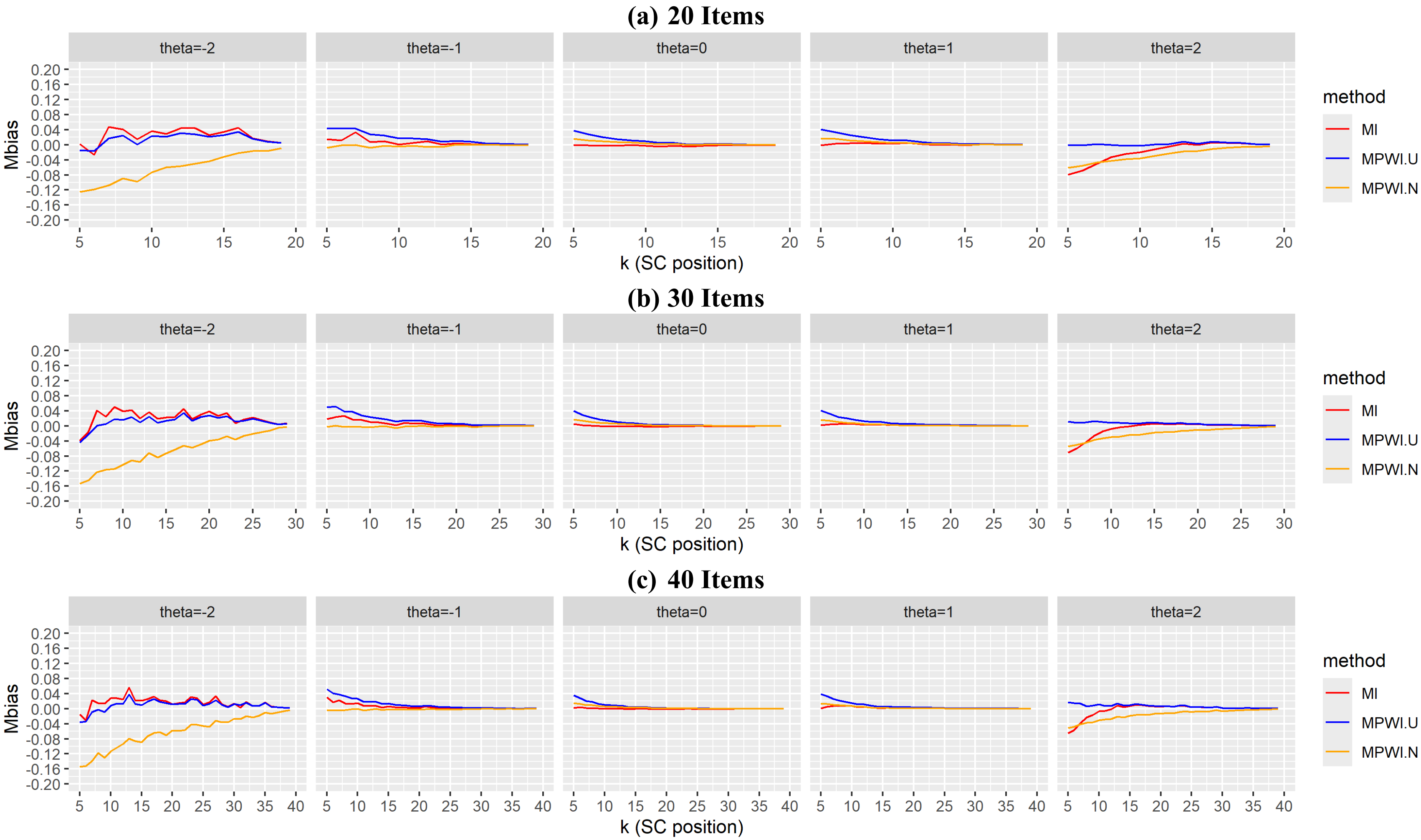


(a) 20 Items

(b) 30 Items

(c) 40 Items

Figure E8. MB of the FSEM estimation procedure under the low-information dichotomous bank

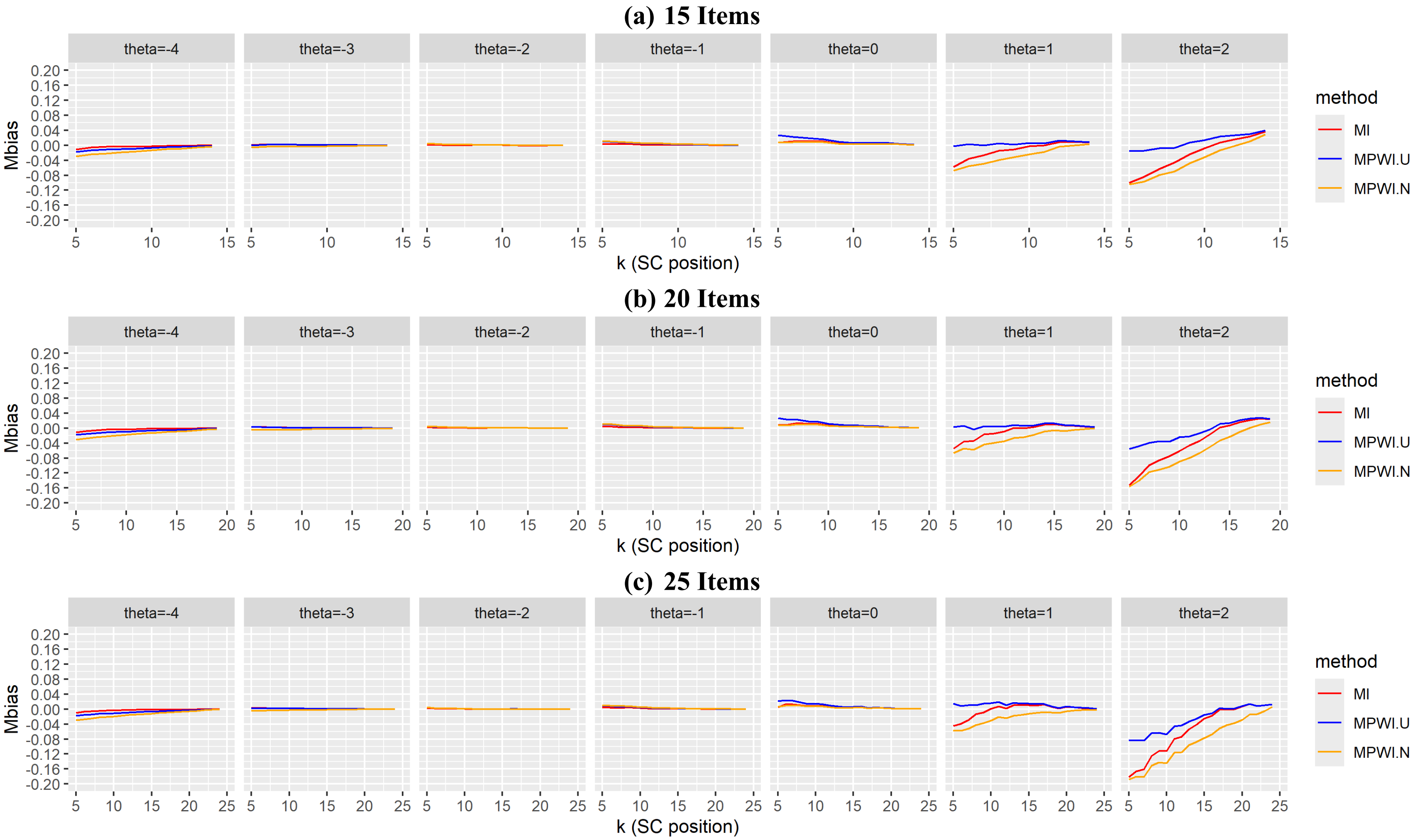


(a) 20 Items

(b) 30 Items

(c) 40 Items

Figure E9. MB of the FSEM estimation procedure under the low-information polytomous bank



(a) 15 Items

(b) 20 Items

(c) 25 Items

Figure E10. Positive predictive values of the SC procedure under the high-information dichotomous banks

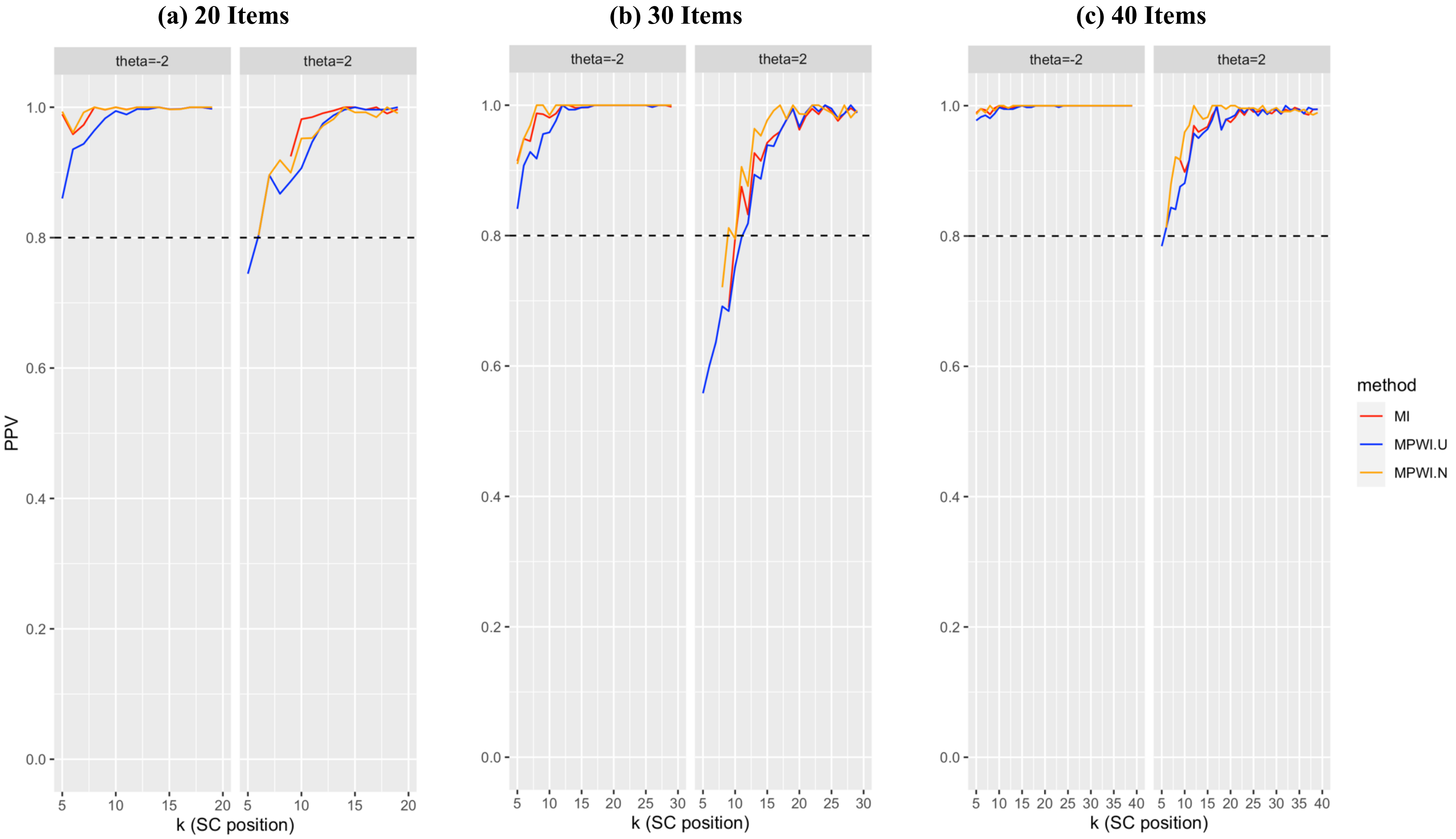

(a) 20 Items **(b) 30 Items** **(c) 40 Items**

Figure E11. Positive predictive values of the SC procedure under the low-information dichotomous banks

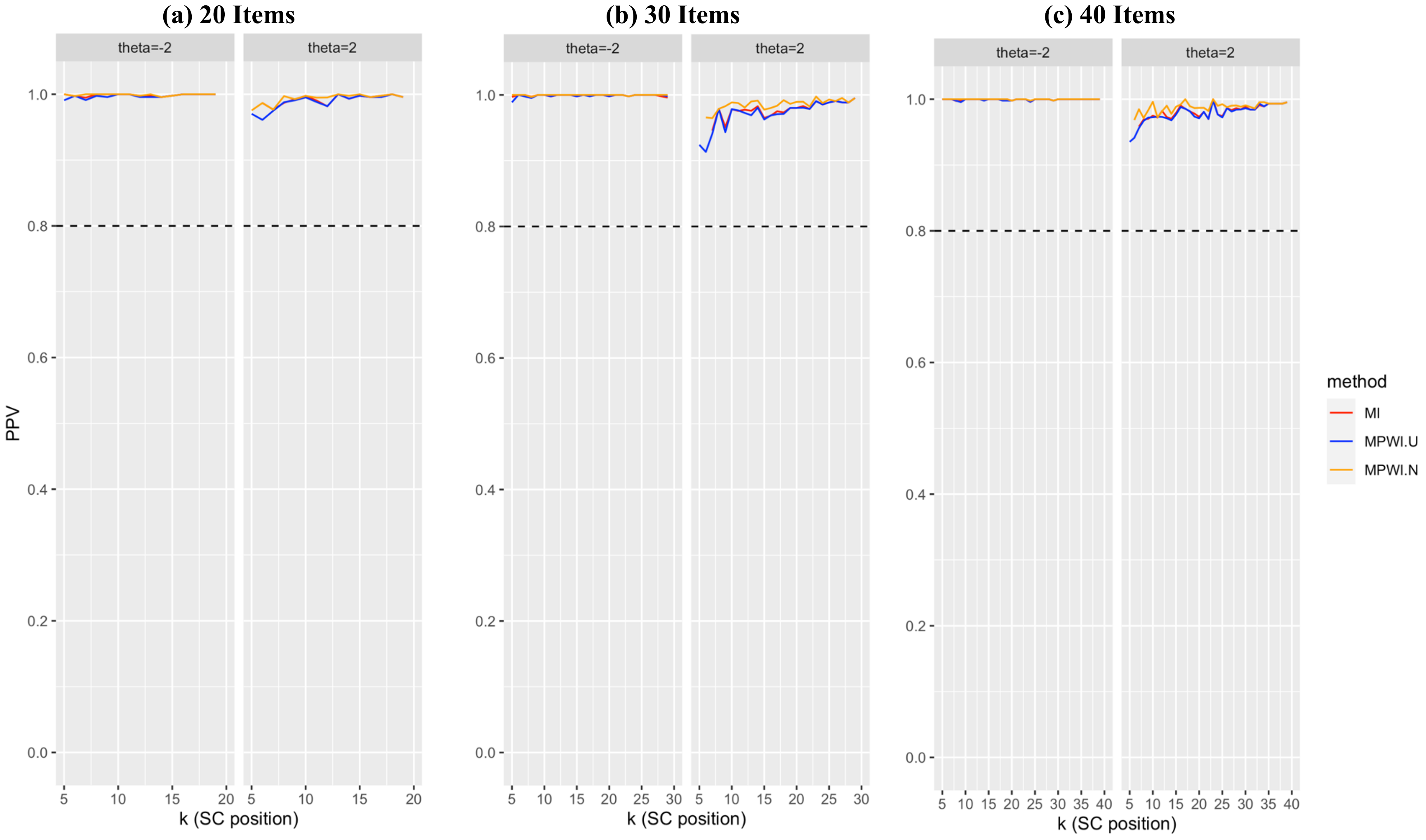
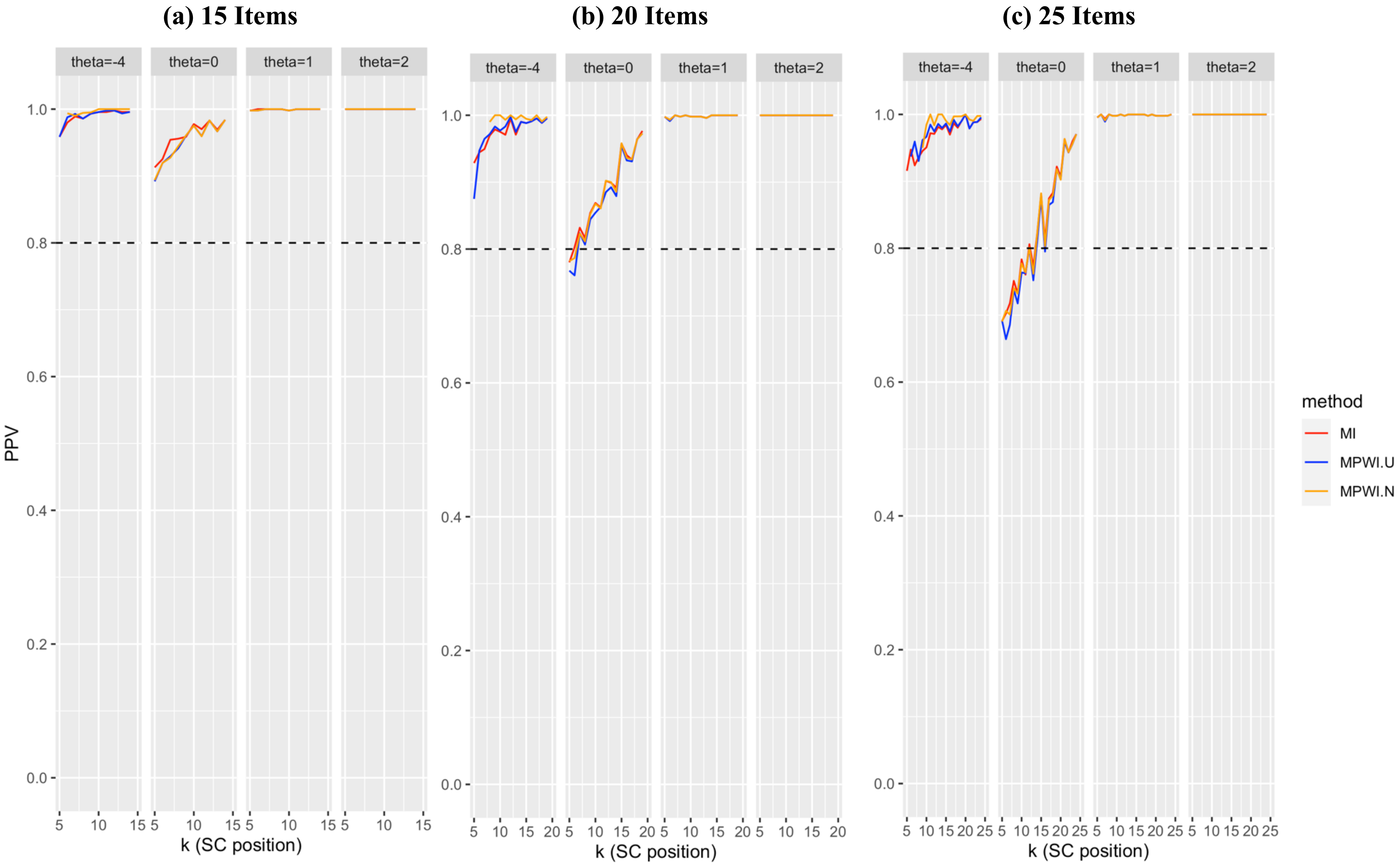


Figure E12. PPVs of the SC procedure under the low-information polytomous bank



(a) 15 Items

(b) 20 Items

(c) 25 Items

Figure E13. True positive rates of the SC procedure under the high-information dichotomous banks

(a) 20 Items **(b) 30 Items** **(c) 40 Items**

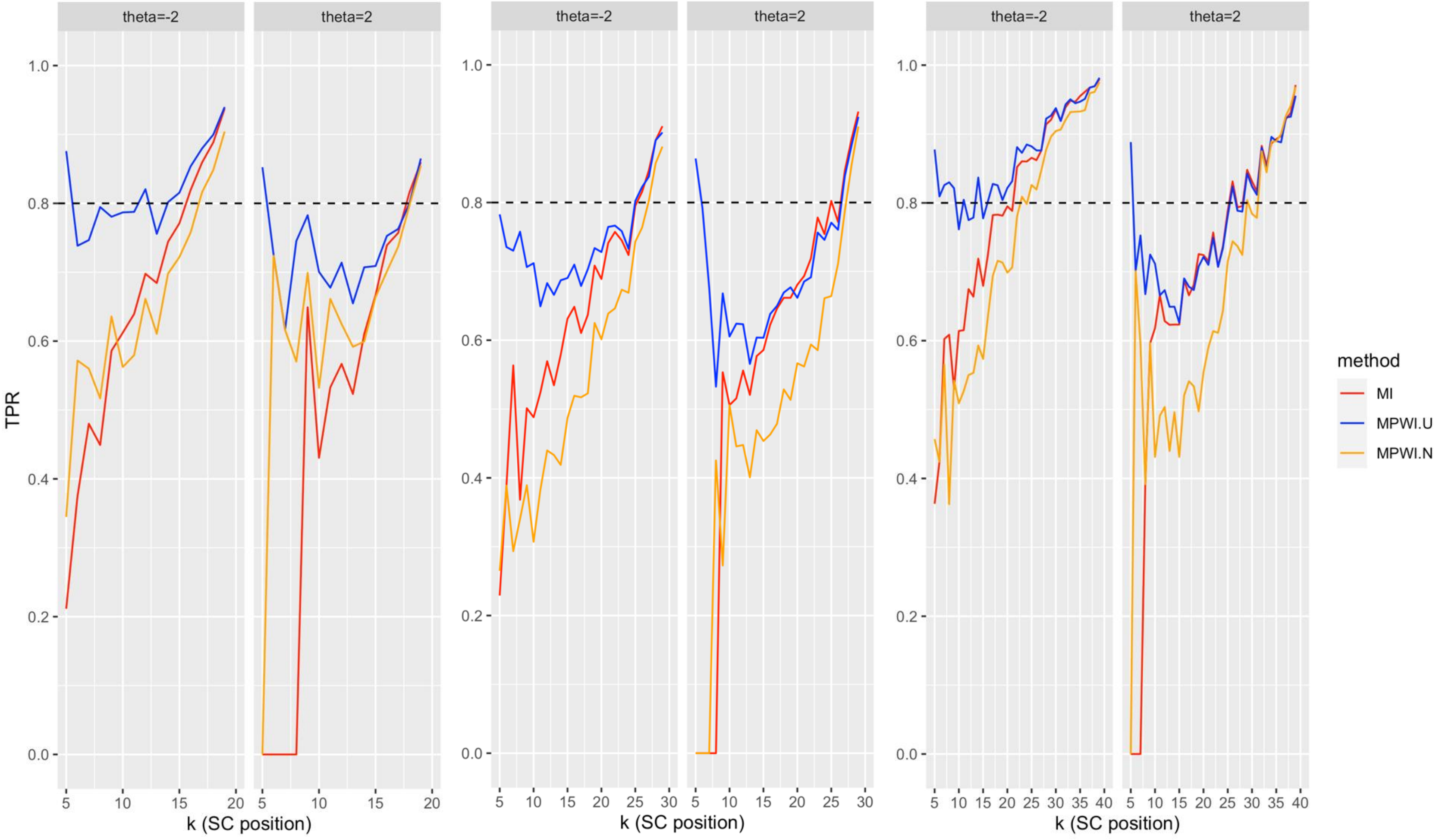


Figure E14. True positive rates of the SC procedure under the low-information dichotomous banks

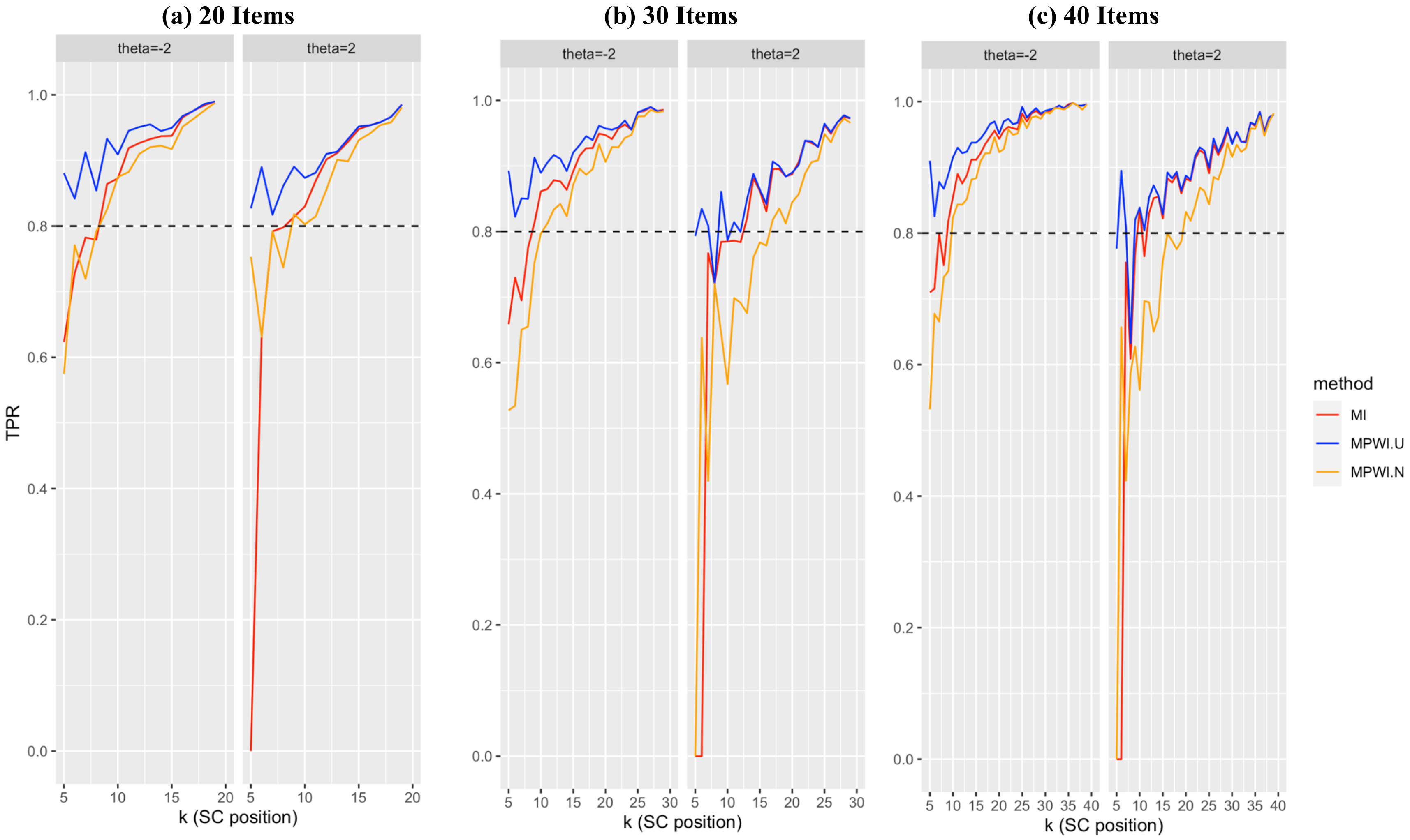

(a) 20 Items **(b) 30 Items** **(c) 40 Items**

Figure E15. True positive rates of the stochastic curtailment procedure under the low-information polytomous bank

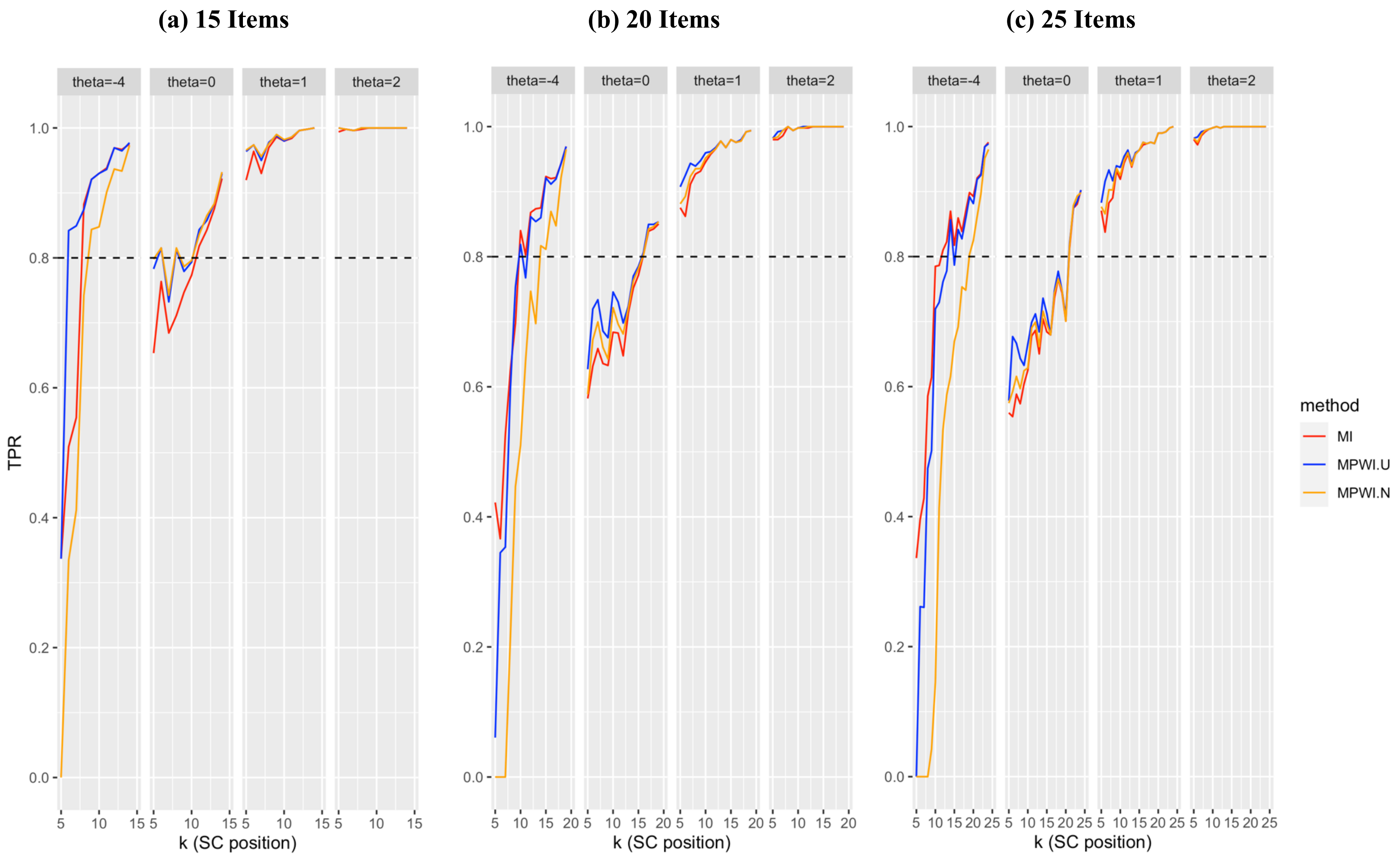

(a) 15 Items **(b) 20 Items** **(c) 25 Items**

Figure E16. False positive rates of the SC procedure under the high-information dichotomous banks

(a) 20 Items **(b) 30 Items** **(c) 40 Items**

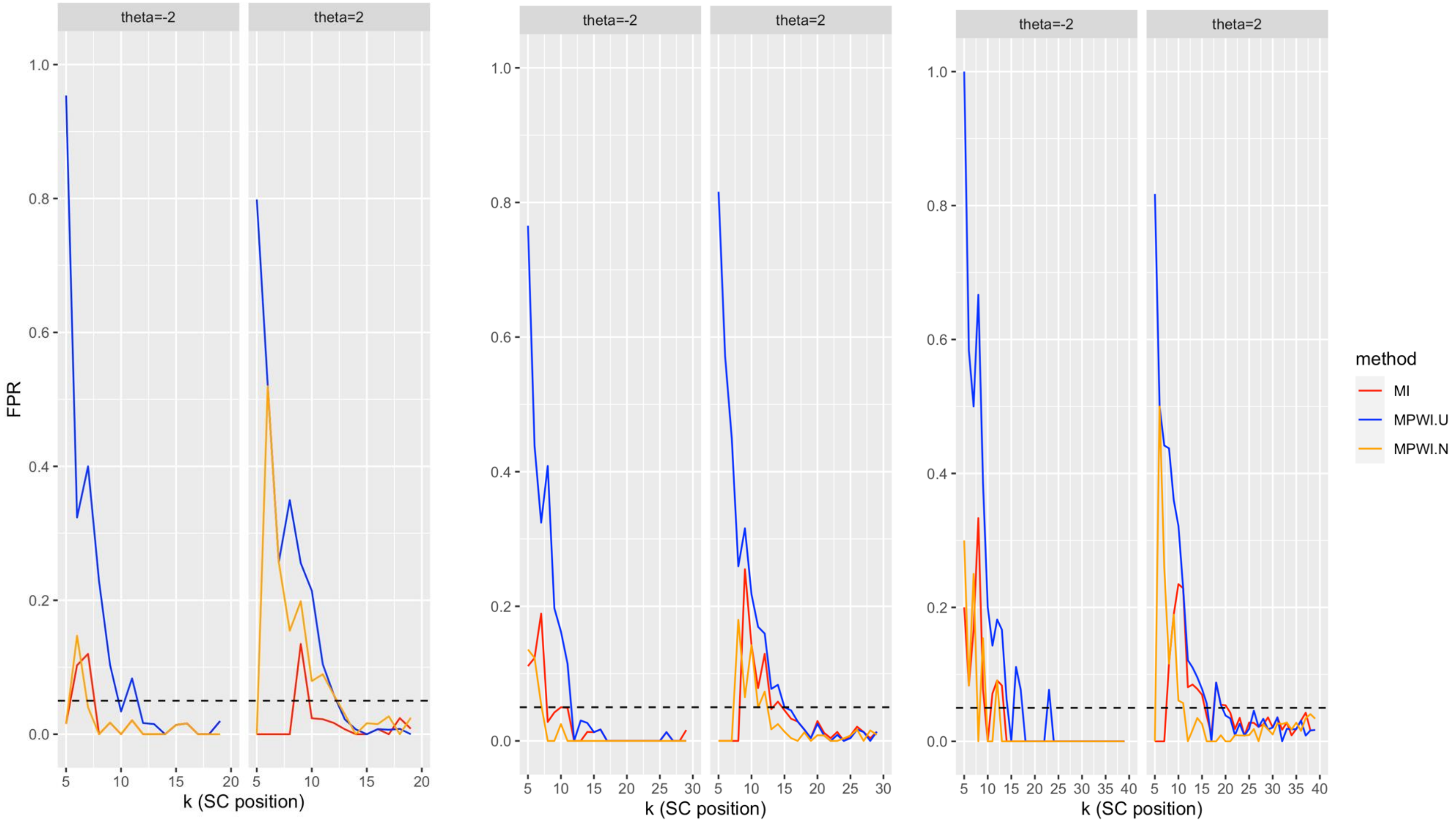


Figure E17. False positive rates of the SC procedure under the low-information dichotomous banks

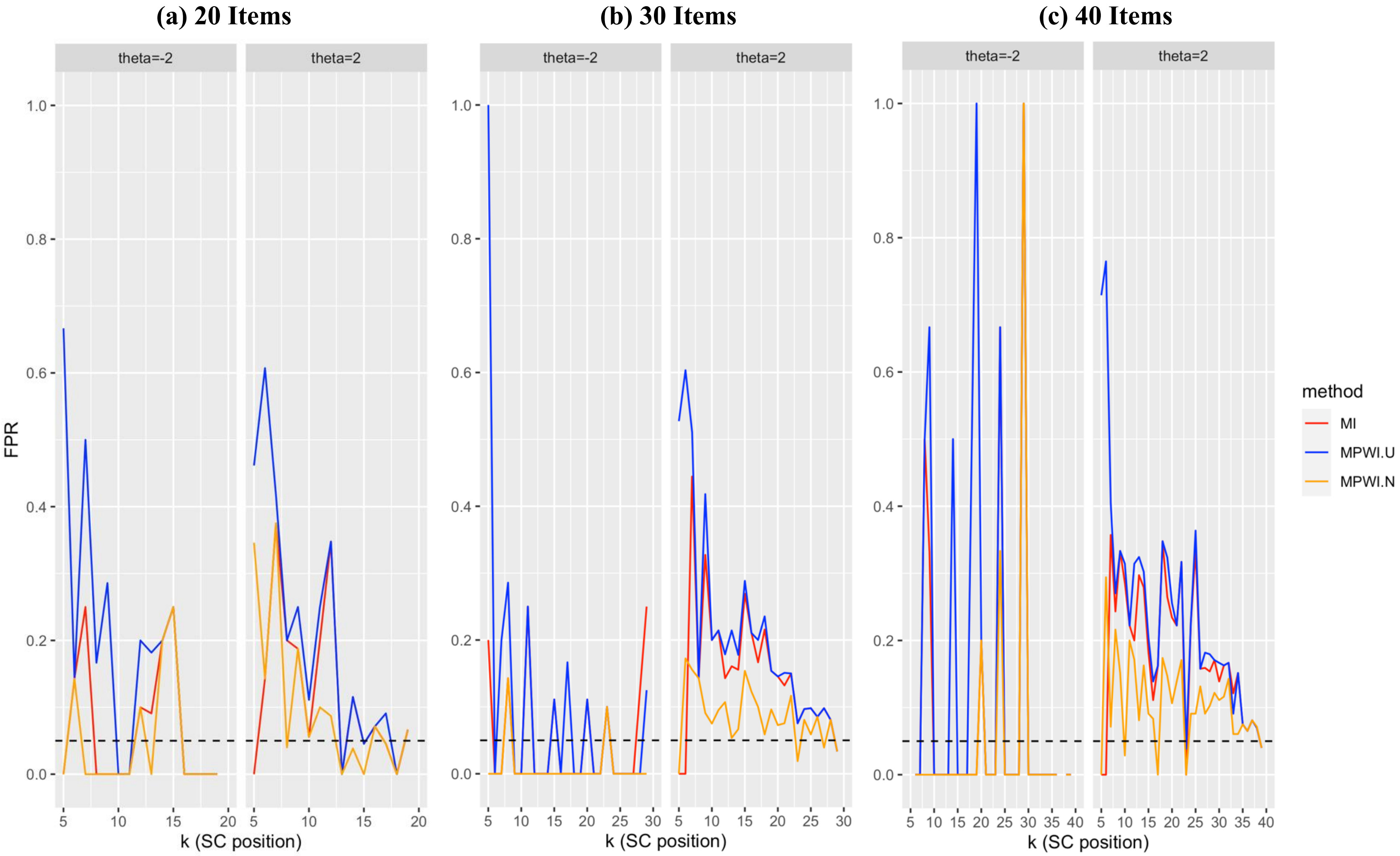
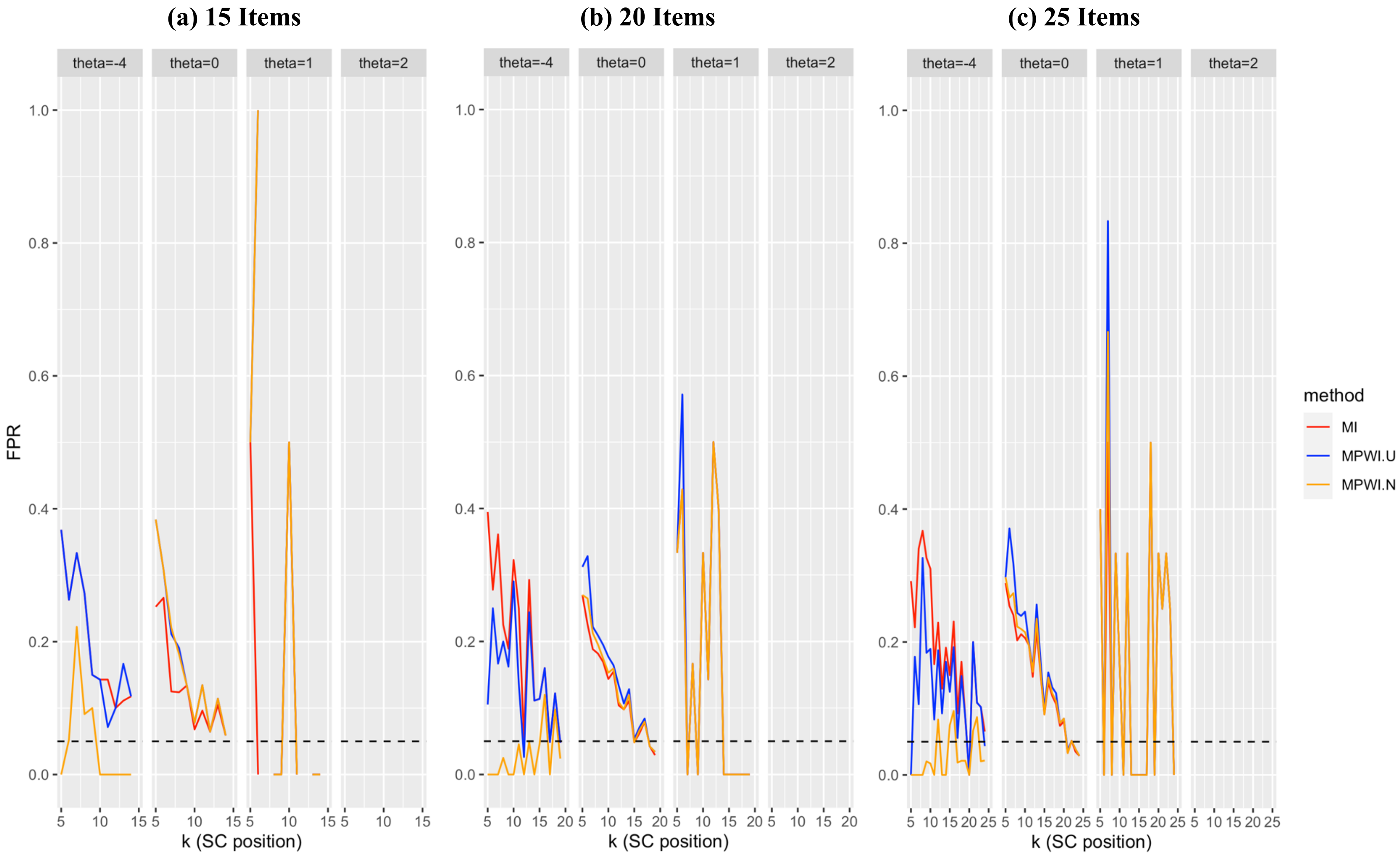


Figure E18. False positive rates of the SC procedure under the low-information polytomous bank



Appendix F

Supplementary Tables

Table F1. Reduction in average test length under the high-information dichotomous bank

Item bank and test length	$\theta = -2$			$\theta = 2$		
	Reduction in ATL	Reduction in ATL (%)	% simulees SCed	Reduction in ATL	Reduction in ATL (%)	% simulees SCed
HI, 20 items	1.9	54.5	89.0	1.9	54.4	86.6
HI, 30 items	18.8	62.7	92.4	17.0	56.7	87.0
HI, 40 items	28.7	71.7	96.8	26.8	66.9	88.6

Table F2. Reduction in average test length under the low-information dichotomous bank

Item bank and test length	$\theta = -2$			$\theta = 2$		
	Reduction in ATL	Reduction in ATL (%)	% simulees SCed	Reduction in ATL	Reduction in ATL (%)	% simulees SCed
LI, 20 items	13.5	67.3	97.8	13.4	67.1	98.0
LI, 30 items	21.9	73.0	98.0	2.3	67.6	95.2
LI, 40 items	32.7	81.7	99.6	29.6	73.9	96.4

Table F3. Reduction in average test length under the low-information polytomous bank

Item bank and test length	$\theta = -4$			$\theta = 0$		
	Reduction in ATL	Reduction in ATL	% simulees SCed	Reduction in ATL	Reduction in ATL	% simulees SCed
LI, 15 items	9.8	65.3%	99.6%	9.1	6.7%	98.2%
LI, 20 items	12.8	64.0%	99.6%	11.5	57.7%	96.6%
LI, 25 items	17.6	7.4%	98.8%	16.6	66.4%	99.0%

Item bank and test length	$\theta = 1$			$\theta = 2$		
	Reduction in ATL	Reduction in ATL	% simulees SCed	Reduction in ATL	Reduction in ATL	% simulees SCed
LI, 15 items	9.8	65.5%	99.8%	1.0	66.6%	10.0%
LI, 20 items	12.9	64.7%	10.0%	13.0	65.0%	10.0%
LI, 25 items	17.9	71.6%	10.0%	18.0	72.0%	10.0%

Appendix G

Supplementary Analyses Under Balanced Conditions

To address concerns regarding the influence of class imbalance on model performance metrics, supplementary analyses were conducted using conditions that spanned the full θ continuum, with particular focus on more balanced scenarios. Specifically, the SC procedure's performance was evaluated across all θ values (−2, −1, 0, 1, 2) for the 20-item dichotomous banks under both low- and high-information conditions. This allowed evaluation of the stability and validity of performance indicators such as positive predictive value (PPV), true positive rate (TPR), and false positive rate (FPR) beyond the extreme θ cases emphasized in the main analysis.

Class Imbalance and PPVs

Class imbalance refers to the predominance of one case type, whether low-precision or high-precision, within the population. As discussed in the main text, PPV was positively correlated with the base rate of low-precision (positive) cases. In extreme θ conditions, the proportion of low-precision cases was naturally high, which inflated PPV estimates. In contrast, at more central θ levels, the base rate of low-precision cases declined sharply, reducing the potential ceiling for PPV. In both scenarios, PPVs rose steadily throughout the test, starting at moderate levels (approximately 40% and 60%, respectively, using MPWI-U as an example) and surpassing 80% by the 10th item (Figures G1 and G7). These results suggested that even under reduced base rates (Figures G2 and G8), SC-based forecasts remained informative, demonstrating "genuine" model performance rather than purely reflecting class imbalance.

TPRs and FPRs

At $\theta = -1$, the TPRs and FPRs provided additional insight into model utility under more balanced conditions. The base rate of low-precision cases at this θ level was 30% in the high-information bank and 50% in the low-information bank (Figures G4 and G10). For TPRs, MPWI-U consistently outperformed the other methods, achieving detection rates between 55% and 85% in both banks as the test progressed (Figures G3 and G9). This indicated that the method correctly identified a substantial proportion of low-precision cases even when they were not highly prevalent. The other two projection-based methods performed less reliably but showed improvement as the test proceeded.

FPRs followed a complementary pattern. At θ = −1, the base rates of negative (high-precision) cases were 70% in the high-information bank and 50% in the low-information bank (Figures G6 and G11). MPWI-U exhibited relatively high FPRs early in the test (approximately 40–50%) but rapidly declined to below 5% after administering 10–12 items (Figures G5 and G12). The other methods started at lower FPRs (~10%) and also dropped below 5% quickly. These findings suggested that the SC procedure—particularly MPWI-U—balanced TPRs and FPRs more effectively as the test accumulated information, even in moderately imbalanced settings.

Summary

These supplementary findings supported the robustness of the SC procedure under more balanced testing conditions and confirmed that its performance was not solely driven by class imbalance. While class prevalence undoubtedly influenced metrics like PPV and TPR, the observed patterns—especially the convergence to high PPV and low FPR after a moderate number of items—indicated meaningful classification utility. Further work could explore the interaction between item bank characteristics, θ distributions, and base rates to optimize SC implementation across diverse CAT contexts.

Figure G1. PPVs of the SC procedure for the 20-item high-information dichotomous bank

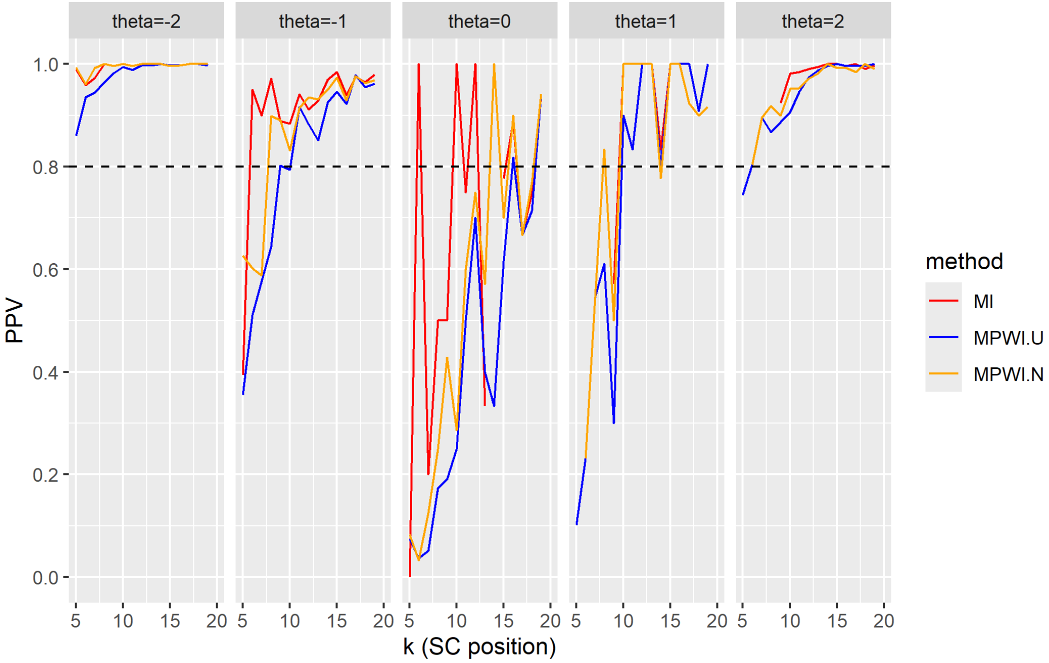


Figure G2. Composition of SC's positive alarms by true positives and false positives for the 20-item high-information dichotomous bank

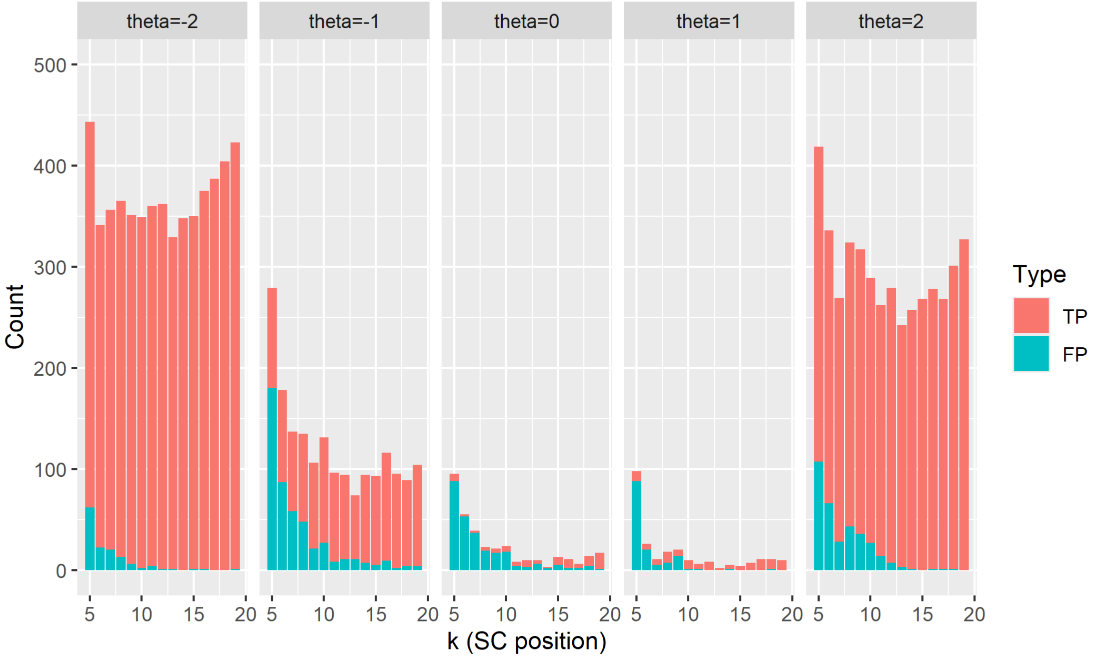


Figure G3. TPRs of the SC procedure for the 20-item high-information dichotomous bank

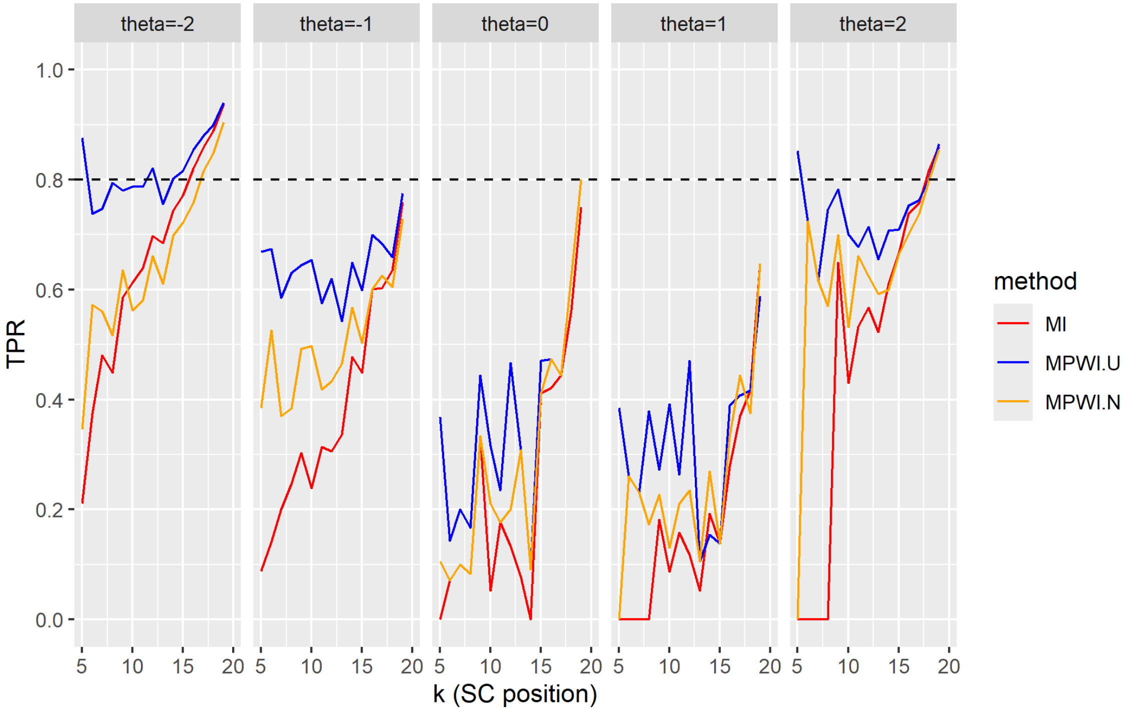


Figure G4. Composition of population positive cases by true positives and false negatives for the 20-item low-information dichotomous bank

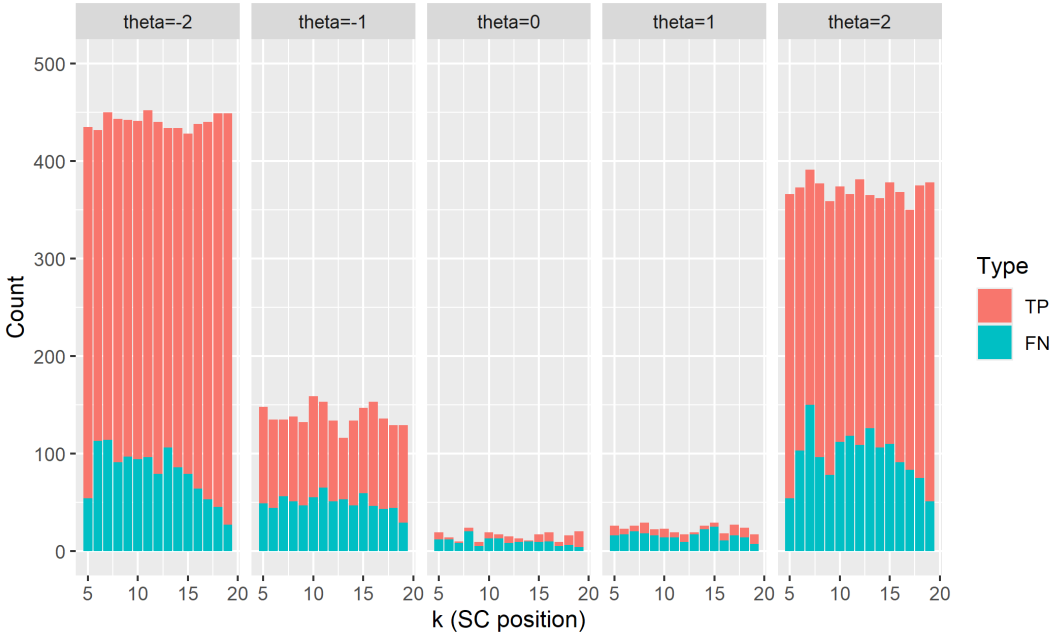


Figure G5. FPRs of the SC procedure for the20-item high-information dichotomous bank

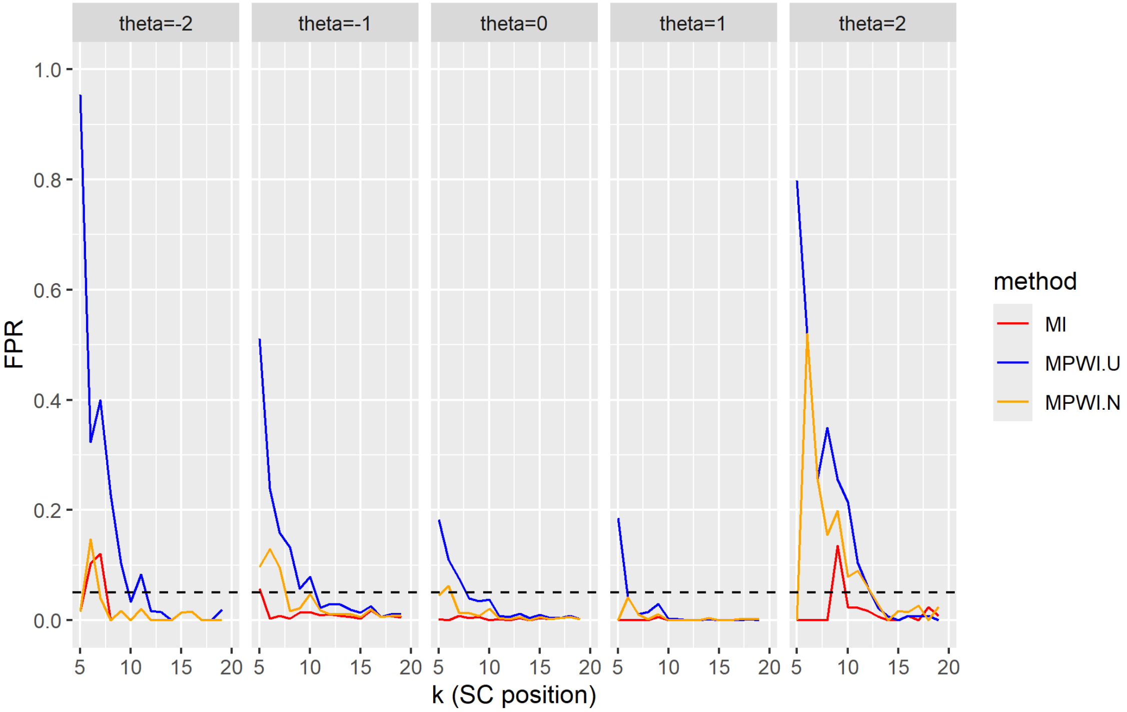


Figure G6. Composition of population negative cases by false positives (FP) and true negatives (TN) for the 20-item high-information dichotomous bank

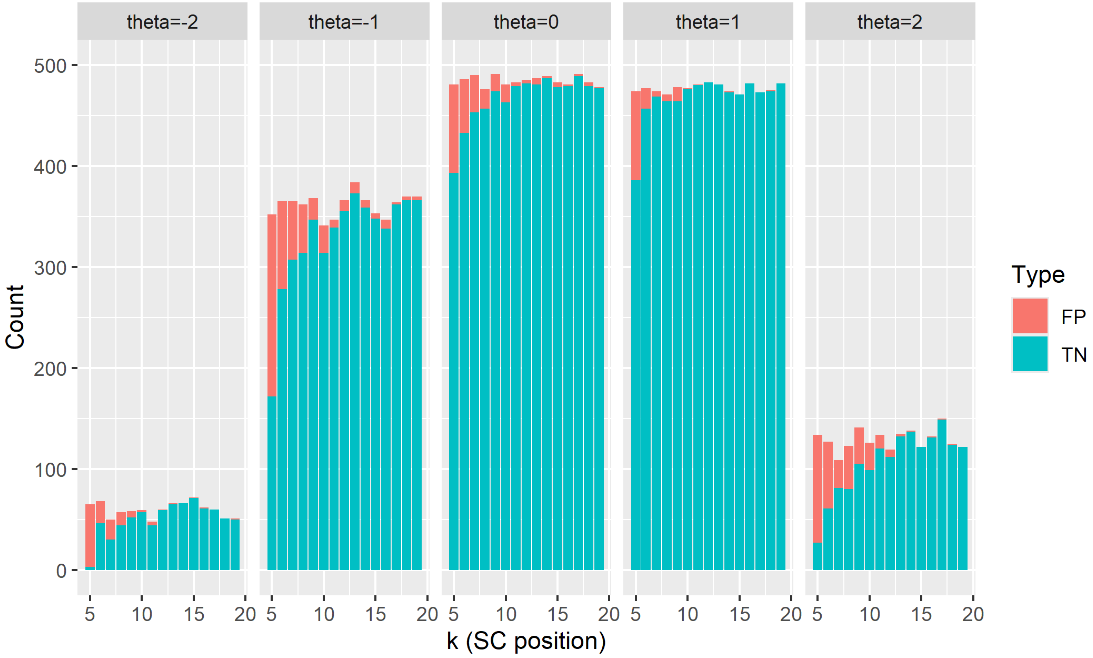


Figure G7. PPVs of the SC procedure for the 20-item low-information dichotomous bank

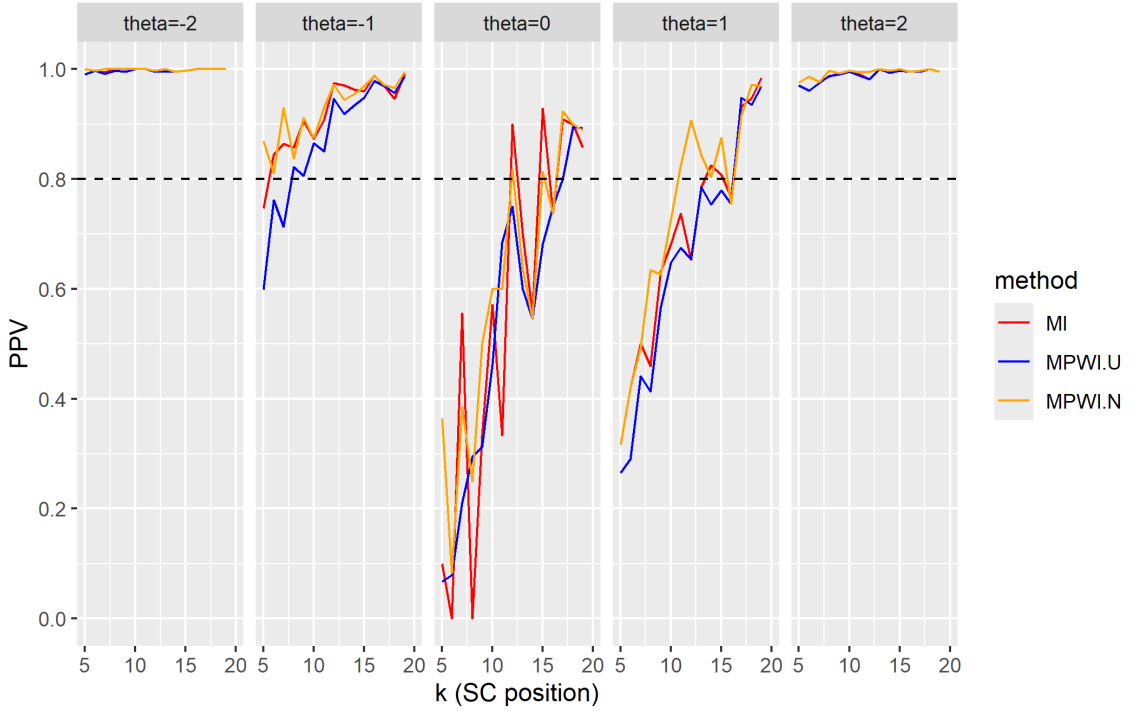


Figure G8. Composition of SC's positive alarms by true positives (TP) and false positives (FP) for the 20-item low-information dichotomous bank

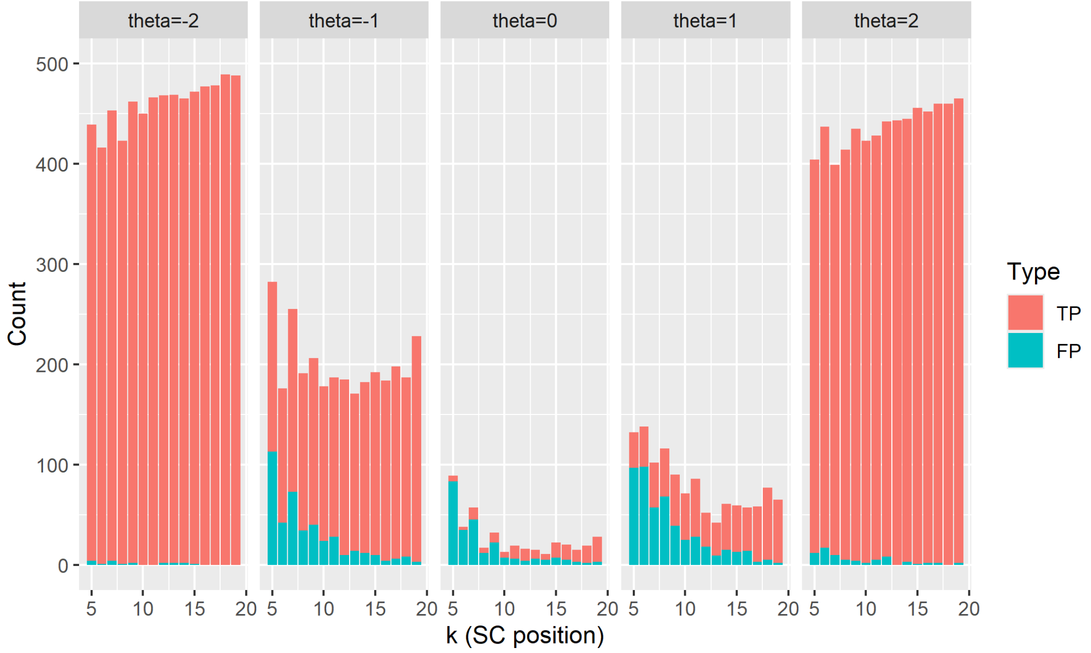


Figure G9. TPRs of the SC procedure for the 20-item low-information dichotomous bank

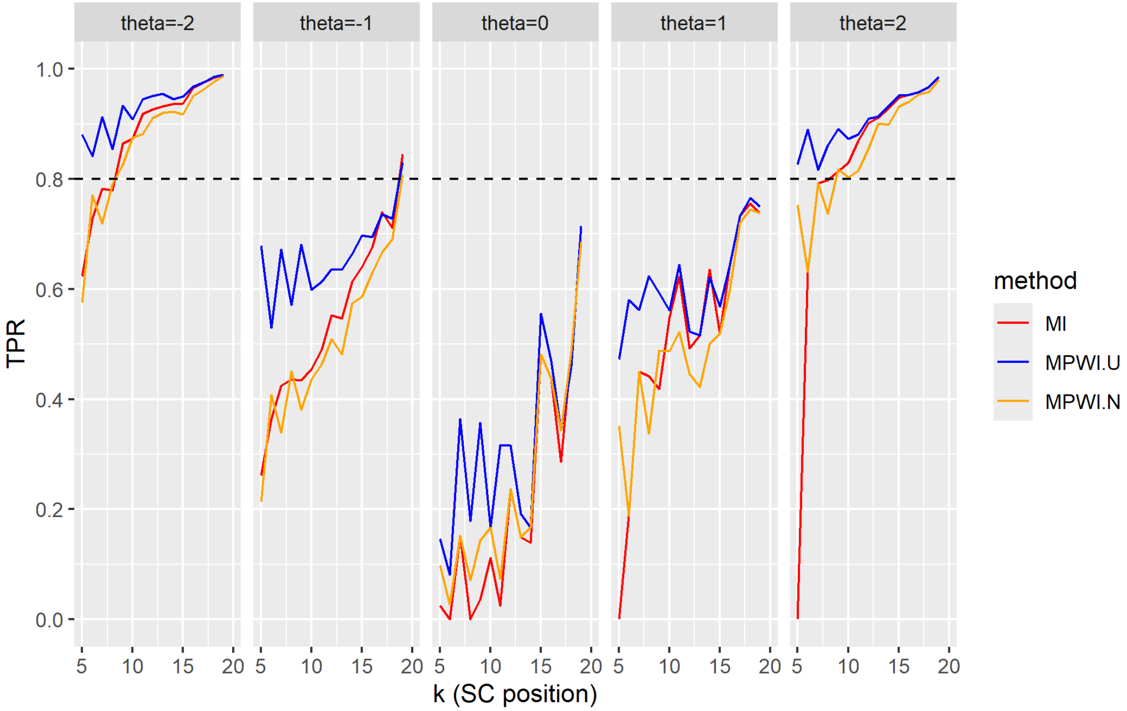


Figure G10. Composition of population positive cases by true positives (TP) and false negatives (FN) for the 20-item low-information dichotomous bank

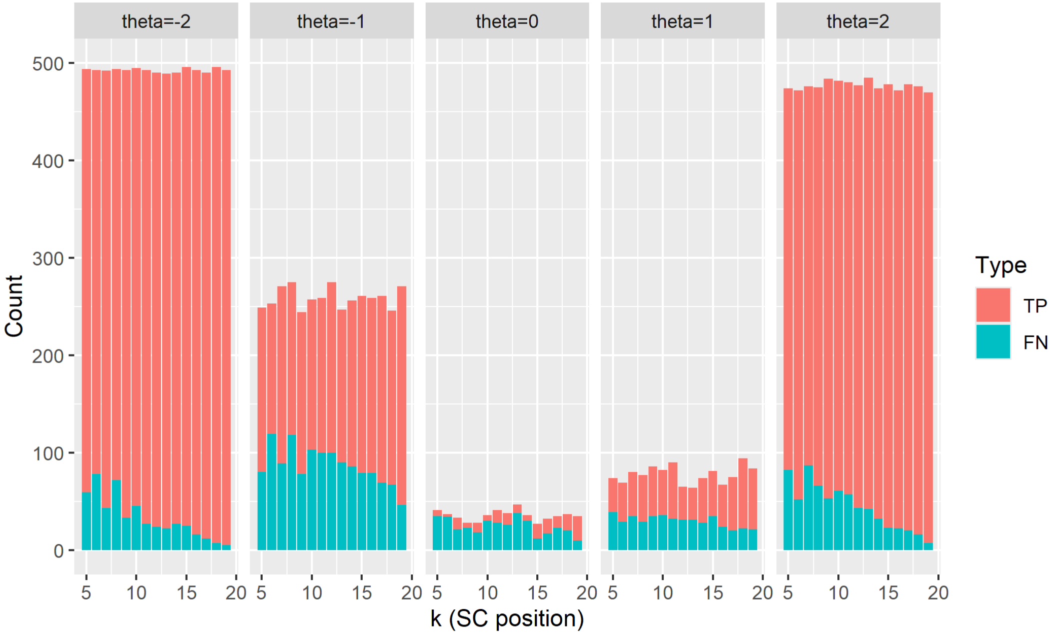


Figure G11. FPRs of the SC procedure for the 20-item low-information dichotomous bank

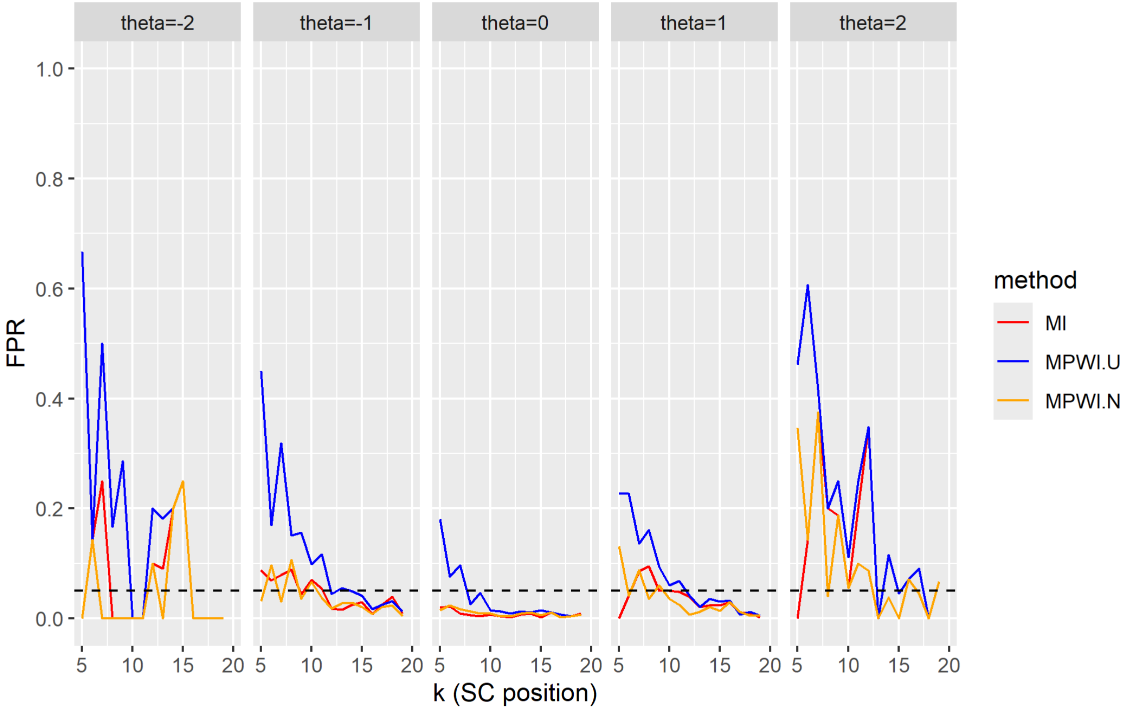
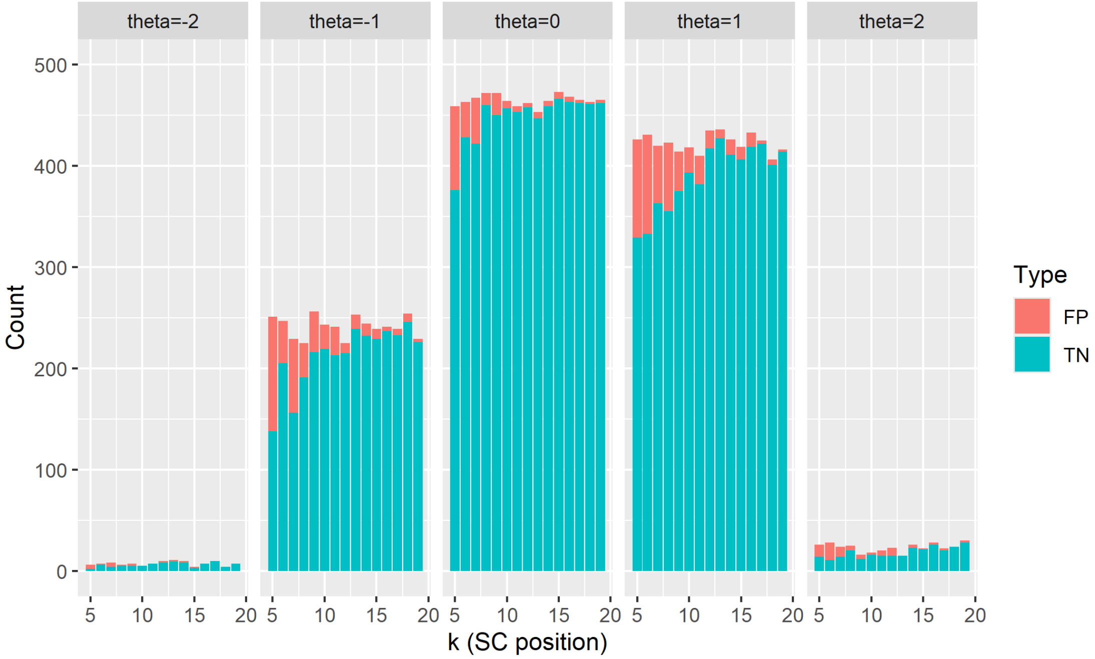


Figure G12. Composition of population negative cases by false positives (FP) and true negatives (TN) for the 20-item low-information dichotomous bank


theta=-2
theta=-1
theta=0
theta=1
theta=2
Count
k (SC position)
Type
FP
TN