ETS GRE

Assess Ability. Predict Performance.

ETS GRE® Board Research Report

ETS GRE®–16-03

ETS Research Report No. RR–16-22

Investigating Robustness of Item Response Theory Proficiency Estimators to Atypical Response Behaviors Under Two-Stage Multistage Testing

Sooyeon Kim

Tim Moses

June 2016

The report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

****

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board reports do no necessarily represent official Graduate Record Examinations Board position or policy.

The Graduate Record Examinations and ETS are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

****

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

GRE-ETS

PO Box 6000

Princeton, NJ 08541-6000

USA

To obtain more information about GRE
programs and services, use one of the following:
Phone: 1-866-473-4373
(U.S., U.S. Territories*, and Canada)
1-609-771-7670
(all other locations)
Web site: www.gre.org

*America Samoa, Guam, Puerto Rico, and US Virgin Islands

RESEARCH REPORT

Investigating Robustness of Item Response Theory Proficiency Estimators to Atypical Response Behaviors Under Two-Stage Multistage Testing

Sooyeon Kim[1] & Tim Moses[1,2]

1 Educational Testing Service, Princeton, NJ
2 Present address: The College Board, New York, NY

The purpose of this study is to evaluate the extent to which item response theory (IRT) proficiency estimation methods are robust to the presence of aberrant responses under the *GRE*® General Test multistage adaptive testing (MST) design. To that end, a wide range of atypical response behaviors affecting as much as 10% of the test items were simulated using a generic GRE 2-stage MST and the 2-parameter logistic (2PL) IRT model. As expected, some differences were found among the 5 estimators in terms of the recovery of the true theta ability; for example, Bayesian estimators had lower error variance and their estimates were regressed to the mean. Once the IRT theta estimates were scaled onto a comparable reporting score scale, however, it was found that all the estimation methods investigated, including the one currently used to score GRE MSTs, were equally robust under the simulated conditions.

Keywords Multistage testing; proficiency estimator; atypical response; item response theory; Bayesian estimator; *GRE*® General Test

doi:10.1002/ets2.12111

In a multiple-choice test, examinees' responses could be classified into the following three types: (a) responses reflecting examinees' true ability; (b) correct responses made through lucky guesses or preknowledge (i.e., item exposure); and (c) incorrect responses derived from anxiety, carelessness, or distraction (Y. Yen, Ho, Laio, Chen, & Kuo, 2012). The latter two types of atypical responses might cause error in proficiency estimation because they would not reflect the examinees' actual knowledge and assumed item response patterns given their true knowledge levels. Meijer (1996) described various factors that can cause an examinee's score on a test to be spuriously high or spuriously low. Those factors are cheating (e.g., copying from another examinee), careless responses, guessing, creative responses, and random responses. Guessing and cheating are well-known atypical behaviors that may artificially inflate proficiency estimates. Some atypical behaviors (e.g., alignment error on a separate answer sheet or copying a neighbor's answers) are obsolete in the present computer mode environment, whereas some other factors (e.g., preknowledge from the Internet) emerge due to technological advancements.

By design, the atypical responses may lead to larger estimation error in adaptive testing than in traditional linear testing (Y. Yen et al., 2012). Various topics related to the examinees' aberrant responses (e.g., person-fit statistics, item selection strategy) have been investigated under the computerized adaptive test (CAT) context (see Ho & Li, 2013; Karabatsos, 2003; Meijer, 1994, 2005; Sijtsma & Meijer, 2001). Chang and Ying (2009) found that when responses to earlier test items were misfit, it would cause the CAT item selection procedure to select items that were not appropriate for the examinee. Misfit responses were more detrimental for high-level examinees than for low-level examinees in terms of total number of dichotomous items required to recover their actual ability levels under CAT (Guyer & Weiss, 2009; Rulison & Loken, 2009). According to their simulation studies, CAT with the conventional 3-parameter logistic item response theory (3PL-IRT) model was unable to recover the (high-level) examinees' true abilities from misfit-as-incorrect responses (perhaps due to their careless behaviors or unfamiliarity to the computer testing environment) within a test length of 45–50 items. This means that the high-ability examinees who respond incorrectly to the first two items in the CAT would not obtain an unbiased ability estimate even after 50 items were administered. Early aberrant responses cannot be discounted with a retrospective evaluation of the entire response vector, as the early answers provide the only information with which to continue the CAT and select future items. Given a sufficient test length (e.g., more than 30 items), however, CAT recovered

Corresponding author: S. Kim, E-mail: skim@ets.org

the (low-level) examinees' true proficiency levels from misfit-as-correct-responses. This result indicates that successful guessing early in the CAT may not result in biased ability estimates unless test length is insufficient.

It is generally assumed that adaptive testing may be more vulnerable to the atypical responses than is traditional linear testing. This statement may not be always true, however, because adaptive testing exists in various formats and performs differently depending upon its design structure. Like CAT, multistage testing (MST) is also adaptive. But its designs vary substantially as a function of numbers of stages, numbers of modules, or numbers of items in modules. Because MST differs substantially from CAT in terms of its adaptive algorithm and design structures, it is difficult to generalize the findings derived from CAT directly to the MST context. The impact of the examinees' atypical responses on their proficiency estimates would differ depending upon the design structure of adaptive testing. It is expected that MST would be more robust than CAT to atypical responses but more vulnerable than fixed-format (linear) tests. Any conclusion for the impact of atypical responses under the MST context cannot be made, however, because research related to this topic is lacking in the literature.

Multistage Testing

CAT and MST use different adaptive algorithms. Under the CAT design, adapting to an examinee's ability occurs at the item level to improve precision and efficiency of measurement, whereas under the MST design, adapting to an examinee's ability occurs between item sets (modules) based on cumulative performance on previous item sets. An MST design consists of a small number of separate modules, and each module can be assembled to meet a set of specifications such as item content and item difficulty. Examinees receive preassembled item sets determined by their performances at previous stages. MST includes only a small number of decision points—in some cases, only one. In particular, the two-stage MST is actually a sequence of two conventional linear tests, with the first test scored before administration of the second test and used to assign each examinee to the appropriate difficulty level of the second test.

The choice of MST design configurations and psychometric characteristics of MST assembly is influenced by various factors, such as test score use (certification or admission), item security, item pool capacity, and administration environments. Adding stages and modules within stages can produce tremendous practical complexity without adding psychometric benefits for the final forms (Jodoin, Zenisky, & Hambleton, 2006; Luecht & Nungester, 1998; Luecht, Nungester, & Hadidi, 1996; Wang, Fluegge, & Luecht, 2012). A recent study (Wang et al., 2012) showed that complex and simple MST designs performed equally well when the item bank consisted of high-quality items targeting key ability regions. The authors of that study recommended the use of simple MST configurations, because those designs can be used with a modestly sized item bank.

The determination of how to allocate items to the Stage 1 routing section and to subsequent sections also depends on the test's purpose and the item pool's quality. H. Kim and Plake (1993) found that the statistical characteristics of the first-stage (routing) module had a major influence on the complete test's measurement precision compared to the later stages' modules. However, more recent studies using a two-stage MST (S. Kim & Moses, 2014; S. Kim, Moses, & Yoo, 2015a, 2015b) showed that the numbers of items at Stage 1 and at Stage 2 were about equally important. When a sufficient number of items is administered at the subsequent stage, the number of items at Stage 1 would not be more important than the number of items at subsequent stage. S. Kim et al. (2015b) assembled eight two-stage MST forms as a function of not only the second-stage module differences in difficulty (overlap vs. distinct) but also module length at Stage 1 and Stage 2 (e.g., 25 items in Stage 1 and 15 items in Stage 2 [called 25–15], 20–20, 15–25, and 10–30). In their simulation, the use of distinct second-stage modules offered a benefit by reducing estimation error at the extreme regions of the proficiency scale. The difference among the four module-length conditions was almost negligible in terms of the accuracy of proficiency estimates. For example, although the 10–30 panel led to substantial misclassification at Stage 1 due to a lack of items available for routing (only 10 items), its final proficiency estimates were as accurate as those in other panel types.

MST provides a compromise between fully adaptive testing (e.g., CAT) and conventional linear form testing. This feature has led to interest in its operational use in practice (see Educational Testing Service [ETS], 2011). Since the launch in August 2011, the *GRE*® revised General Test (rGRE) has followed a two-stage MST design that includes one module at Stage 1 and three modules at Stage 2, with 20 items in each module. Each examinee receives a total of 40 items across the two stages, which are separately timed at 30 minutes per stage. The GRE program uses a 2PL-IRT model to calibrate item parameters using the data from the domestic examinees and uses number-correct scoring (often called inverse test characteristic curve [TCC]) to produce examinees' reported scores.

The MST's unique feature also led to various studies (see Luecht & Sireci, 2011, for a summary). Even so, many studies related to MST focused on design issues such as adequate numbers of stages, numbers of modules, or panel assembly (see Jodoin et al., 2006; Luecht & Nungester, 1998; Luecht et al., 1996; Wang et al., 2012). Very little guidance appears in the literature about the effects of the choice of an estimator in practical applications, and the impact of atypical responses on the estimation process. Although a recent study compared the performance of IRT estimation methods under the MST framework (S. Kim et al., 2015a, 2015b), research on the impact of atypical responses under MST is clearly lacking in the literature. This topic is important because the choice of scoring method has a direct influence on the examinees' reported scores or proficiency levels, and the examinees' atypical responses are unavoidable in practice.

Proficiency Estimation Methods

Given that some level of aberrance is unavoidable in actual test data, the question is how to reduce the potential impact of atypical behaviors on examinees' proficiency estimates and ultimately reported scores. The choice of proficiency estimators (i.e., scoring method) represents one approach to answering that question. An examinee's proficiency level can be estimated in a variety of ways when using an IRT model, and an MST panel (i.e., form) can be scored in several ways. Generally well-known IRT proficiency estimation methods are as follows:

1. Test characteristic function with number-correct scoring[1] (TCF)
2. Maximum likelihood estimation with item-pattern scoring (MLE)
3. Expected a posteriori with number-correct scoring ($_S$EAP)
4. Expected a posteriori with item-pattern scoring (EAP)
5. Maximum (mode) a posteriori with item-pattern scoring (MAP)

MLE has been used commonly in practice. MLE finds the examinee's proficiency level that maximizes the likelihood of obtaining the examinee's observed test data, given the item parameters and model. In other words, MLEs are the parameter values that maximize the likelihood that the observed data would have been generated. Thus MLE values correspond to the mode of the likelihood function.[2] Desirable properties of the MLE are that it is asymptotically unbiased and that its standard error is related to the information function (Baker, 1992). Drawbacks of MLE, however, include infinite estimates for examinees whose response patterns are either only incorrect (extremely low proficiency) or only correct (extremely high proficiency). To overcome this, the range of values for the estimated proficiency level can be restricted. This remedy is equivalent to estimating the proficiency level using a Bayesian estimator with a uniform prior distribution in the selected range. Bayesian methods such as EAP and MAP do not share this limitation. Under the Bayesian paradigm, the posterior distribution of the proficiency levels (i.e., θ) is defined as the product of the likelihood function and the prior ability distribution. Bayesian methods incorporate information about the prior distribution to approximate the posterior distribution of latent proficiency (Bock & Mislevy, 1982). The mean of the posterior distribution is the proficiency estimate under EAP, whereas the mode of the posterior distribution is the proficiency estimate under MAP (W. M. Yen & Fitzpatrick, 2006). The choice of a reasonable prior distribution for the proficiency level is key to Bayesian estimators. The most common prior distribution is the standard normal distribution, $N(0, 1)$, and the estimated proficiency levels are shrunk toward the prior mean value. Bayesian estimators generally lead to biased estimates, but their overall errors tend to be relatively small due to shrinking to the mean (EAP) or mode (MAP). The shrinkage is expected to be more pronounced under number-correct scoring, owing to its lower precision, than under item-pattern scoring (Kolen & Tong, 2010). This trend did not emerge, however, when the conditional variances of the item-pattern and number-correct scoring methods were compared (see S. Kim at al., 2015a, 2015b).

Kolen and Tong (2010) described psychometric properties (e.g., reliability, conditional standard errors of measurement [CSEM], and score distributions) of IRT proficiency estimators. Using real data examples, they demonstrated that the choice of estimators can have a significant influence on practical applications of IRT, such as score distribution or the assignment of examinees to proficiency levels. They commented that very little guidance appears in the literature about the effects of choice of estimator in practical applications, although psychometricians are aware of the theory underlying those estimators. A few researchers compared the performance of IRT proficiency estimators and their impact using either real or simulated datasets (Magis, Béland, & Raîche, 2011; Tong & Kolen, 2007; Tong & Kolen, 2010). Recently, S. Kim et al. (2015a, 2015b) compared the performance of seven IRT proficiency methods under the two-stage MST design using simulated datasets. They showed that Bayesian estimators performed better than non-Bayesian ones mainly at the two extreme score regions of the theta scale. Although the difference between item-pattern scoring and number-correct scoring was almost negligible in the portion of the score range where most examinees' scores were located, number-correct scoring produced more stable ability estimates than item-pattern scoring for high-performing examinees. Even so, because the estimation results derived from different proficiency estimators were comparable across the theta region where most examinees would be located, the authors recommended the use of a simpler method (e.g., non-Bayesian number-correct scoring) as a practical choice particularly in the operational setting of the two-stage MST.

Examinees with the same number-correct score but different item response patterns are likely to obtain different proficiency estimates under item-pattern scoring but the same estimates under number-correct (i.e., summed) scoring. There have been some debates regarding the practical benefits of number-correct scoring versus the psychometric benefits of item-pattern scoring. Number-correct scoring is easier for test users to understand than is item-pattern scoring. Certain testing programs prefer item-pattern scoring to number-correct scoring, however, because item-pattern scoring offers more precise estimations than number-correct scoring does. According to Yen and Fitzpatrick (2006), however, when tests include 30 or more items, the inverse of the TCC (a number-correct score) "provides a very accurate MLE of ability for the 3PL model" (p. 137). Based upon the theoretical models underlying the scoring methods, however, it is expected that number-correct scoring may be less sensitive to atypical response behaviors than item-pattern scoring. Also, use of prior distribution information may enhance the accuracy of proficiency estimates under this atypical circumstance.

Purpose

With IRT, the operational scoring process that results in an examinee's reportable score includes essentially two steps: (a) the estimation of examinees' proficiency (on the IRT theta scale) and then (b) converting the proficiency estimate onto the reporting score scale.

The purpose of this study is to investigate the extent to which the alternate proficiency estimators available lead to reported scores that are robust to the presence of atypical responses under two-stage MST with 20 items at each stage. To accomplish this objective, we created a typical GRE two-stage MST panel (i.e., test form) with each item parameterized according to the 2PL-IRT model and simulated examinees' true ability across a wide range of ability levels and their responses. When simulating the examinees' responses on the MST items, we manipulated a particular atypical response type in order to determine which proficiency estimators produce the most accurate results for each type of response deviations.

The first part of our evaluation was to compare the proficiency estimates of each estimator with the examinees' true proficiency levels at the theta scale for each atypical type. Given the conditions in which various atypical responses existed under MST, we compared psychometric effectiveness of number-correct scoring versus item-pattern scoring and examined the potential of prior information to enhance the overall precision of proficiency estimates.

The second part of our evaluation was to compare the estimation results at a hypothetical reporting score level in order to assess direct impact on the examinees' final scores. In this study, however, we focused only on the precision in the final estimation, not on the intermediate estimation at Stage 1, because the previous studies based on both the empirical dataset and simulated dataset showed that the impact of misrouting was minimal for the two-stage MST, which is the same as the one used in the present study (see S. Kim & Moses, 2014; S. Kim et al., 2015b; S. Kim, Robin, & Liu, 2012).

Method

Multistage Test Panel Assembly

Figure 1 illustrates a two-stage MST panel design that was used in this study. At Stage 1 (often called routing), there is only one module; all examinees taking that panel are tested with the same set of items. At Stage 2, there are three modules: low, medium, and high difficulty. Each module at Stage 2 concentrates on a particular level of difficulty to differentiate examinees' abilities within a certain range of proficiency after routing. The items an examinee receives at Stage 2 are determined by the examinee's performance on Stage 1. The term *path* can be used to mean a combination of modules that could possibly be presented to an examinee. There are three paths in the two-stage MST panel, and each path consists of the first-stage module and one of the second-stage modules. The total number of items that each examinee received was 40.

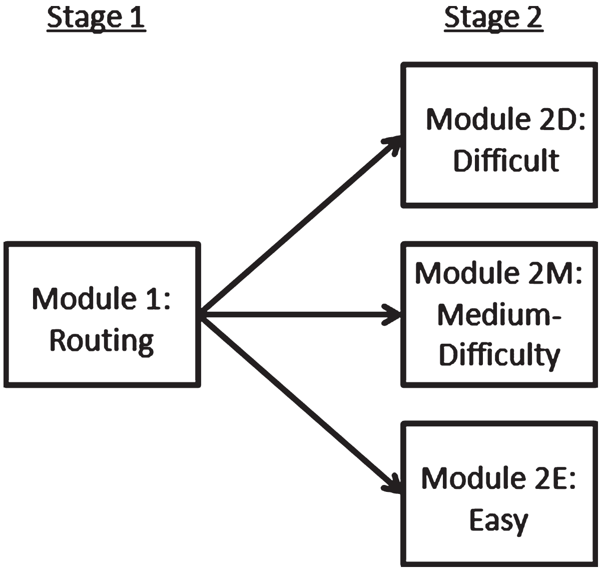

Figure 1 Schematic of a two-stage multistage test (MST).

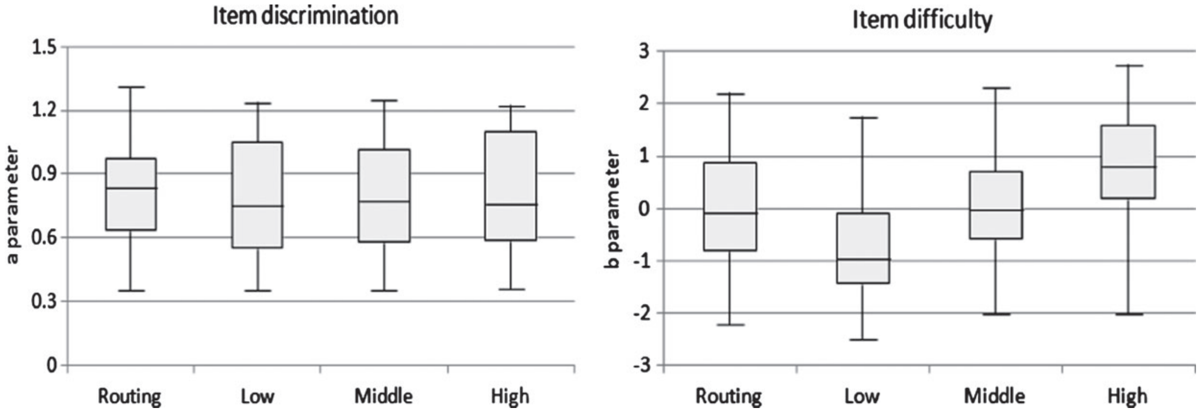

Figure 2 Box-and-whisker plots for each module of the multistage test (MST) panel used in this study.

We assembled the MST panel (i.e., test form) with a 2PL-IRT model based solely on statistical specification (e.g., item difficulty, item discrimination, and number of items). We chose the 2PL-IRT model because large-scale testing programs (e.g., the rGRE and the *TOEFL*® test) use the 2PL model operationally, and the benefit from use of a 3PL rather than a 2PL model appeared to be small in a simple score test of the normal 2PL model (Haberman, 2006). To make the simulated panel realistic, we examined more than 1,000 MST panels' statistical properties that had been administered in actual operational settings. The average of item difficulty parameters was set to be 0.00 for routing, −0.75 for low, 0.00 for middle, and +0.75 for high. The averaged item discrimination parameters were set at 0.80 in all modules. Item parameters for each of four modules were generated using the Microsoft Excel random number generator function.

Figure 2 graphically presents the distributions of item difficulty as well as item discrimination for each of the four modules using box-and-whisker plots. In each module, items were ordered by difficulty level. Thus the easiest item appeared first and the hardest item appeared in the end of the test. We used the TCC and the test information function (TIF) to assess the extent to which each module and each path were reasonable.

Figure 3 displays four plots that present the TIFs and TCCs for the MST panel used in this study. Two plots in the first column indicate the TIF and TCC for each module, and two plots in the second column indicate the TIF and TCC for each path. Perhaps the TIF is more useful under item-pattern scoring, whereas the TCC is more useful under number-correct scoring in an actual assembly setting. Because we used both scoring methods in this simulation, we considered both the TIF and the TCC in assessing the reasonableness of the simulated MST panel.

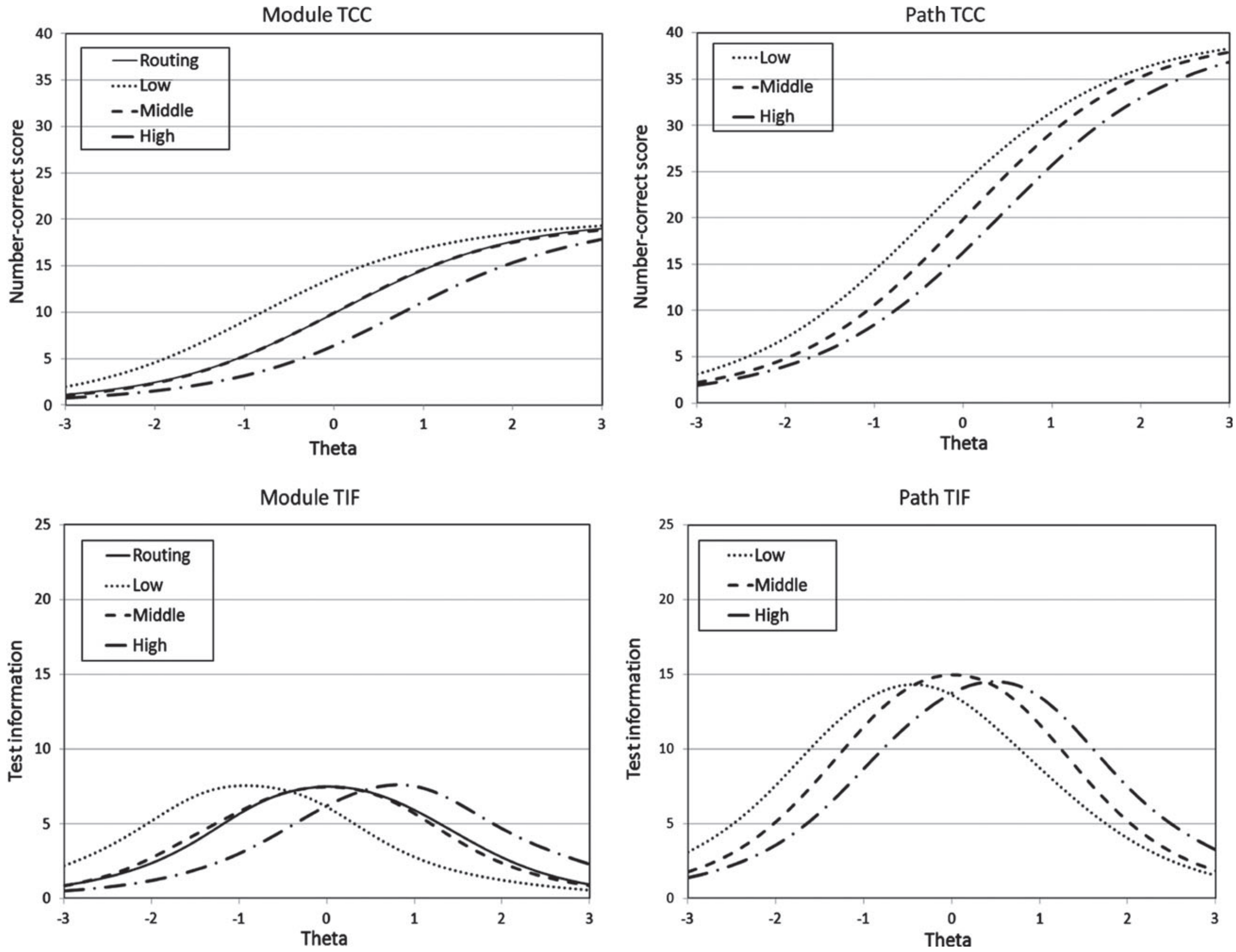

Figure 3 Test information functions (TIF) and test characteristic curves (TCC) for each module of the multistage test (MST) panel used in this study.

Table 1 Characteristics of the Item Response Theory (IRT) Proficiency Estimation Methods

	Scoring	
Prior distribution	Number correct	Item pattern
No (non-Bayesian)	TCF	MLE
Yes (Bayesian)	$_{S}$EAP	EAP, MAP

Notes. TCF = test characteristic function; MLE = maximum likelihood estimation; $_{S}$EAP = expected a posteriori with number-correct scoring; EAP = expected a posteriori; MAP = maximum (mode) a posteriori.

Proficiency Estimators

Table 1 summarizes five IRT proficiency estimation methods that were investigated in this study. Those methods can be classified as a function of scoring method and use of prior information. TCF and $_{S}$EAP use number-correct scoring, whereas MLE, EAP, and MAP use item-pattern scoring. TCF and MLE are non-Bayesian estimators, whereas $_{S}$EAP, EAP, and MAP are Bayesian estimators. In this study, we used the standard normal distribution as a prior for the three Bayesian estimation methods. The specific formulas for each estimation method are as follows.

In the 2PL-IRT model, the probability of correct response is modeled as a function of a latent proficiency level (θ) and item parameters:

$$P_h(\theta) = P\left(X_h = 1|\theta, a_h, b_h\right) = \frac{1}{1 + \exp\left[-1.702a_h\left(\theta - b_h\right)\right]}, \tag{1}$$

where a_h is the item discrimination (or slope) parameter, and b_h is the item difficulty (or threshold) parameter. Estimates of the examinees' proficiency levels $(\widehat{\theta})$ were obtained by treating the item difficulties (b) and discriminations (a) as known parameters in the H items' 2PL IRT models in Equation 1 and using each of the five proficiency estimators.

Using TCF, examinees' proficiency estimates were defined as those that produced summed scores for the 2PL-IRT model (i.e., true scores under the IRT model) that closely approximated examinees' summed total test scores, such that

$$\sum_h u_h \approx \sum_h P_h\left(a_h, b_h, \theta\right) \quad \text{or} \quad X \approx \tau(\theta), \tag{2}$$

where u_h indicates the sum of correct responses to the H items, X indicates the summed score, and τ indicates true summed score (Kolen & Tong, 2010). Using MLE, examinees' proficiency estimates were defined as the θ values that maximized the likelihood of examinees' observed patterns of responses to the H items $(U = (u_1, u_2, \ldots, u_H)^t)$, as shown in Equation 3:

$$L(U, \theta) = \prod_h \left\{ P_h\left(a_h, b_h, \theta\right)^{u_h} \left[1 - P_h\left(a_h, b_h, \theta\right)\right]^{(1-u_h)} \right\}. \tag{3}$$

The maximization was accomplished using the Newton–Raphson algorithm (Baker & Kim, 2004).[3]

In Bayesian estimation, multiplying the likelihood function by a prior distribution yields the posterior distribution. Using $_{\text{S}}$EAP, examinees' proficiency estimates were defined as average proficiency values based on the likelihood of their summed total test scores $(X = \sum_h u_h)$ and an assumed standard normal proficiency distribution:

$$_{\text{S}}\text{EAP} = E(\theta|X) = \frac{\int_\theta \theta L(X, \theta)\, g(\theta)\, d\theta}{\int_\theta L(X, \theta)\, g(\theta)\, d\theta}, \tag{4}$$

where $L(X, \theta)$ was computed using the recursion algorithm and the assumed $g(\theta)$ distribution was approximated with a discrete distribution containing 41 θ values and quadrature points. Using EAP, examinees' proficiency estimates were defined as average values based on the likelihood of their response patterns and an assumed standard normal proficiency distribution:

$$\text{EAP} = E(\theta|U) = \frac{\int_\theta \theta L(U, \theta)\, g(\theta)\, d\theta}{\int_\theta L(U, \theta)\, g(\theta)\, d\theta}. \tag{5}$$

As in operational practice, the standard normal distribution for θ and the integration of this distribution were accomplished by approximating the continuous proficiency distribution, $g(\theta)$, using a discrete distribution with 41 θ values and quadrature points (Baker & Kim, 2004). EAP and MAP follow essentially the same equation. Using MAP, however, examinees' proficiency estimates were defined as those that maximized the likelihood of examinees' response patterns for an assumed standard normal proficiency distribution. Finding the maximum value was accomplished using the Newton–Raphson algorithm (Baker & Kim, 2004).

Atypical Response Types

Atypical responses can occur for numerous reasons, although different forms of atypical behaviors may result in the same kind of item response pattern. Seven atypical responses were used to introduce response patterns that can appear in practice. Table 2 summarizes the description of each atypical response, the direction of estimation bias due to aberrance, and the method used to generate each atypical response. We imposed an atypical response type on the examinees' item responses generated from the IRT model at each examinee's proficiency level. Although we adopted the method used by Karabatsos (2003) in order to generate atypical examinees, the choice of b parameter values associated with lucky guessing, careless responding, or creative responding was rather arbitrary. As for the item exposure, speededness, and peculiar subgroup conditions, we utilized some indirect information obtained from the actual data to make the degree of aberrance more realistic.

Table 2 Seven Atypical Response Types Considered in This Study

Atypical type	Description	Proficiency estimates artificially	How to generate atypical examinees
Random responding	Examinees randomly choose an option of some multiple-choice items on the test.	Increased/decreased	Impute random responses (i.e., no functional relationship with examinees' proficiency level) to four items (i.e., 10% of the test) over the theta region of [−3.0, +3.0] to represent examinees who blindly respond.
Guessing (lucky guessing)	Examinees guess the correct answers to some items for which they do not know the correct answers.	Increased	Assign a .20 probability of a correct response to difficult items (items with b ≥ .5) for the examinees located in the theta range of [−3.0, +0.0].
Preknowledge due to item exposure (cheating)	Examinees unfairly obtain the correct answers on test items that they are unable to answer correctly. Those items no longer discriminate effectively.	Increased	Impute correct responses to some items for the examinees located within the theta range of [−3.0, +3.0]. About 10% of the test; four items randomly selected from the 40 items (e.g., two items are from each stage)
Careless responding	Examinees answer certain items incorrectly for which they are able to answer correctly.	Decreased	Assign a .5 probability of an incorrect response to easy item (b < = −0.5) for the examinees located in the theta region of [+0.5, +3.0].
Creative responding	Examinees answer certain easy items incorrectly because they interpret easy items in a unique and creative manner.	Decreased	Impute incorrect responses to easy items (items with b < = −1.2) for the examinees located in the theta region of [+1.5, +3.0].
No responding to the final items	Examinees cannot reach the final items because the test is speeded.	Decreased	At stage 1, about 5 – 15% of the items are speeded as a function of the theta level (e.g., 15% [−3.0 to −1.0]; 10% [−1.0 to +1.0]; 5% [+1.0 to +3.0]). At stage 2, 10% of the items are speeded in each module.
Peculiar subgroup (more able in ability in this study)	Subgroup's item responses differ from the typical response pattern.	Increased/Decreased	Generate item responses of peculiar subgroup using fake item parameters (e.g., 15% of the examinees at each theta point; subtract 0.3 from the true *b* parameters to make all items less difficult; subtract 0.1 from the true *a* parameters to make all items less discriminating)

Procedure

We compared the five estimation methods in terms of their proficiency estimation accuracy using simulated response data derived from a two-stage MST panel in a situation where each of the seven atypical responses was present. We simulated 1,000 examinees at each of 41 quadrature points on a theta scale ranging from −3.0 (*minimum*) to +3.0 (*maximum*), with an interval of 0.15 ($N = 41{,}000$). The simulation procedure for each response under each of the seven atypical types consisted of the following steps:

1. Generate the simulated examinee's response to each item in the Stage 1 module.
2. Compute the examinee's proficiency (i.e., theta) on Stage 1 using each of the five proficiency estimators, and assign the examinee to the Stage 2 module given the provisional proficiency.

3. Generate the examinee's response to each item in the Stage 2 module.
4. Compute the examinee's proficiency on Stage 1 and Stage 2 combined, using each of the five proficiency estimators, to determine the examinee's estimated proficiency theta.
5. Compute the examinee's reported score, using each of the five theta estimates and their corresponding scoring conversions.

Based upon the defined population intervals method (Luecht, Brumfield, & Breithaupt, 2006),[4] we selected two cut scores for routing simulated examinees to a second-stage module in a manner that would result in approximately 30%, 40%, and 30% of the examinees taking high, middle, and low paths, respectively, as the distribution of the simulated examinees' proficiency estimates. Therefore, the two cut scores associated with the 30th and 70th percentiles of the cumulative distribution of theta may have differed slightly depending upon the estimators.

Given that all simulated examinees' true proficiency levels (i.e., thetas) are known, their estimated thetas can be compared to their true thetas as a method of evaluating the effectiveness of the five estimators. Based on the difference between estimated and true values, three deviance measures—bias, error, and root mean square error (RMSE)—were calculated at each of the 41 quadrature points using the following formulas.

$$\mathrm{Bias}_{jk} = \sum_{i} \left(\widehat{\theta}_{ijk} - \theta_k \right), \tag{6}$$

$$\mathrm{Error}_{jk} = SD\left(\widehat{\theta}_{ijk} \right), \tag{7}$$

$$\mathrm{RMSE}_{jk} = \sqrt{\mathrm{Bias}_{jk}^2 + \mathrm{Error}_{jk}^2}, \tag{8}$$

for a fixed proficiency group θ_k, where i indicates an examinee, j indicates a proficiency estimator, $\widehat{\theta}_{ijk}$ indicates a proficiency estimate of a particular estimator under a particular atypical condition at a theta point k, and θ_k indicates a true proficiency value at a theta point k. As overall summary measures, root mean squared bias, error, and RMSE were each averaged across all quadrature points, weighting the values at each theta level according to its relative percentage (f_k) under the standard normal distribution.[5] For each proficiency estimator, the resulting statistics were the weighted root mean squared bias, $\sqrt{\sum_k f_k \mathrm{Bias}_{jk}^2}$; the weighted standard error of estimation, $\sqrt{\sum_k f_k \mathrm{Error}_{jk}^2}$; and the weighted RMSE, $\sqrt{\sum_k f_k \mathrm{RMSE}_{jk}^2}$. We used a root mean square averaging procedure to prevent negative bias at one score level from canceling out positive bias at another.

When IRT methods are used, theta estimation is considered an intermediate step to producing the final score on a predetermined reporting scale that maximizes scores' usage and interpretation (Robin, 2014). Score scales other than the IRT-θ scale are often found to be more useful for score reporting. In actual test settings, testing programs report scaled scores to the examinees after applying a particular transformation procedure to the thetas to facilitate score interpretation. The IRT proficiency estimators' differences occurring on the reporting scale will be of practical significance because they have direct impact on the examinees' final scores. In reality, the choice of a scaling procedure will depend on the test's purpose (e.g., certification, admission), the population, test length, and psychometric properties, and so on. To make meaningful comparisons among the five proficiency estimators, we chose a hypothetical score scale that may well reflect an existing large-scale high-stakes testing program. The hypothetical scaled scores are approximately normally distributed with mean of 150 and standard deviation of 8, and ranged from 130 to 170, with an increment of 1, after the truncation of the score scale.

The score conversions associated with their respective proficiency estimate were obtained by specifying the target scaling population distribution (here, set to a normal distribution with a mean equal to 150 and a standard deviation equal to 8.0) and using the nonlinear score transformation procedures described in the book of Kolen and Brennan (2004, pp. 338–340). The procedure for finding the true and each one of the estimation method's scaling conversions consisted of the following steps:

1. Find the population theta probabilities (i.e., weights) at each population theta value.[6]

2. Compute the cumulative probabilities and percentile ranks from Step 1 at each population theta value and at each estimation methods' estimated theta values.
3. Take the percentile rank and solve for the score on the standard normal distribution having the same cumulative distribution function value (i.e., the inverse of the standard normal density function).
4. Obtain the mean and standard deviation for the values obtained in Step 3.
5. Apply the mean and sigma scaling by subtracting the mean from Step 4, multiplying by 8 (standard deviation from Step 4), and adding 150 to get the desired mean and standard deviation.

Consequently, the distribution of the scaled true ability and the distributions of the scaled estimates for each method can be expected to have nearly the same mean, variance, and shape under the no aberrance condition. Thus, the scaled scores obtained with each of the five proficiency estimators were compared to the true scaled score at each true theta value (the conditional results). These results were then aggregated accounting for the population weights, into overall deviance measures, in the same way it was done with theta proficiency estimates.

Results

Theta Scale

For each of the five proficiency estimators, we compared examinees' estimated proficiency levels (i.e., theta values) to their true proficiency levels under each of the seven atypical response conditions and the condition of no atypical responses (i.e., no aberrance). Figures 4–6 represent bias, error, and RMSE conditioned on each of the 41 quadrature points across the theta scale, respectively. In the figures, each plot represents each of the following conditions, respectively: no aberrance, random responses, guessing, preknowledge, careless responses, creative responses, missing responses due to speededness, and a peculiar subgroup's responses. The bias plots display the theta differences (e.g., estimated minus true) associated with each proficiency estimator under each of the atypical conditions. Negative bias indicates underestimation, whereas positive bias indicates overestimation in the estimated proficiency levels. The error plots display conditional standard deviations of the differences, which can be interpreted as empirical estimates of the CSEM, under the atypical response conditions. The RMSE plots display the conditional total error, which is a combination of bias and error. Table 3 presents the summary of the weighted root mean squared bias, error, and RMSE that were averaged across the entire theta scale from −3.0 to +3.0, derived under the eight response conditions.

As shown in Figure 4, the five estimators' biases under the random, preknowledge, and peculiar subgroup conditions were nearly identical to the bias associated with no aberrance in terms of magnitudes and patterns. Consequently, as presented in Table 3, the average weighted root mean squared biases under those atypical conditions were almost indistinguishable from the no aberrance condition, and their biases were consistently smaller compared to the other conditions' biases. Under the no aberrance, random, preknowledge, and peculiar subgroup conditions, the five estimators' biases differed noticeably at the top and bottom regions of the theta scale. MLE produced the least bias, TCF produced the second least bias, and the Bayesian methods produced the largest bias. Bayesian methods yielded large positive bias (i.e., overestimation) at the lower theta region and large negative bias (i.e., underestimation) at the upper theta region. By design, all five estimators produced similar patterns of negative bias across the theta region above 0.5 (under the careless condition) or 1.5 (under the creative condition). The careless condition produced larger biases than did the creative condition, which produced slightly larger biases than did the no aberrance condition. The abrupt changes occurring at either at 0.5 or 1.5 of the theta scale were caused by the aberrant manipulation imposed on those conditions. Under the speededness condition, all methods produced negative bias across the theta region above −2.0. The number of not reached items at Stage 1 differed as a function of the examinees' theta level. For example, at Stage 1, the low-performing examinees missed the last three items out of the 20 items, whereas high-performing examinees missed the last item only, which is the most difficult item in the routing module. At Stage 2, however, it was designed that all examines could not reach the last two items. Although the effect of speededness spread out rather widely across the theta scale, its impact was more salient at the upper region of the theta scale because the not-reached items were difficult ones. As presented in Table 3, the magnitude of the average bias was larger under the careless and speededness conditions than other conditions. Under the guessing condition, all five methods performed very similarly to the no aberrance condition across the upper theta region. However, both TCF and MLE produced large positive bias, as did other Bayesian methods, across the theta region lower than −1.0. Thus the guessing condition produced slightly larger biases than did the no aberrance condition.

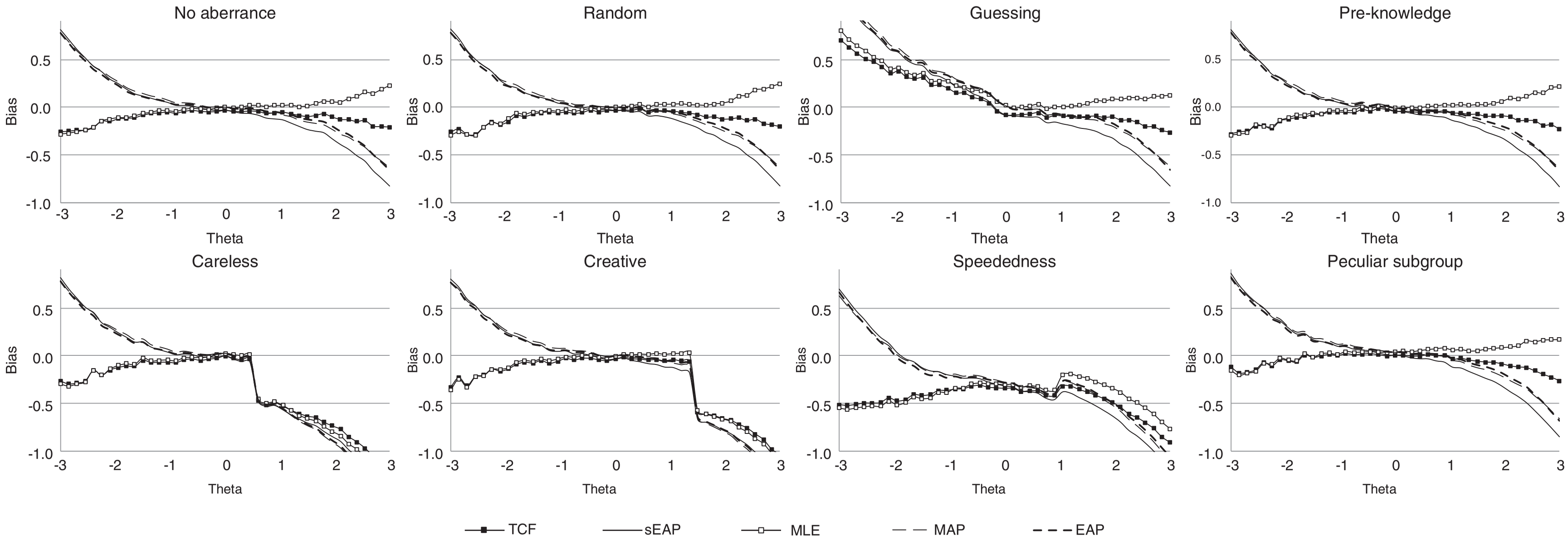

Figure 4 Conditional bias of each estimation method on the theta scale.

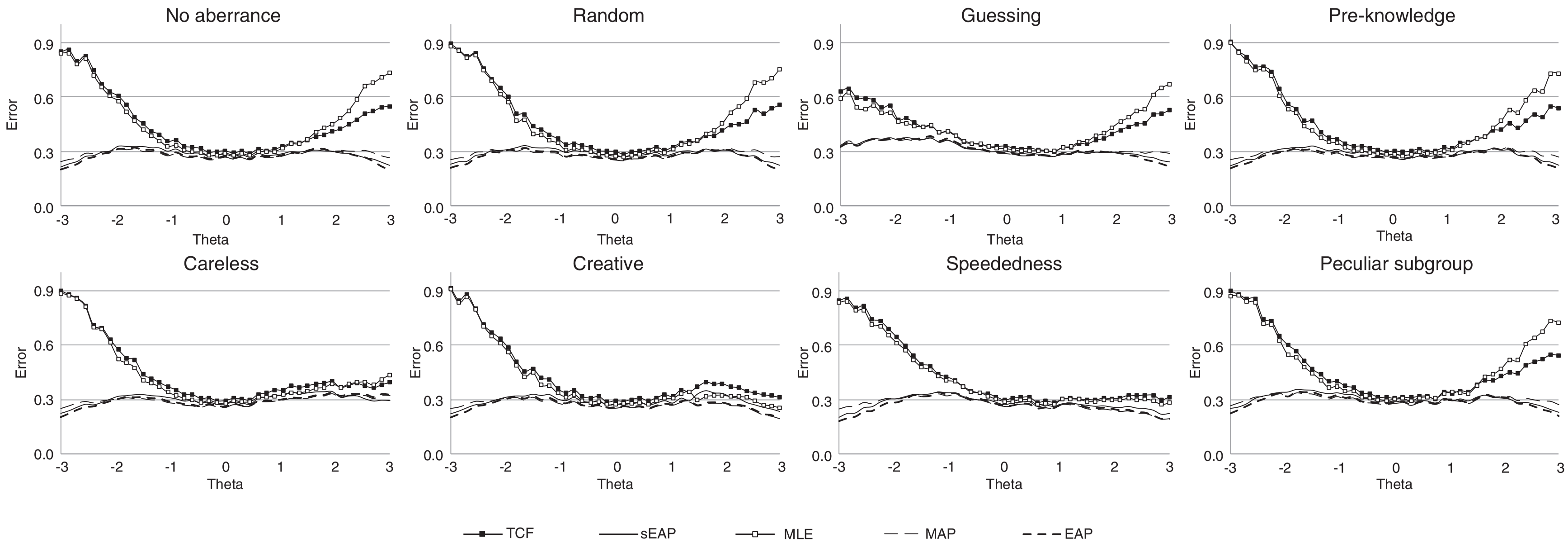

Figure 5 Conditional error of each estimation method on the theta scale.

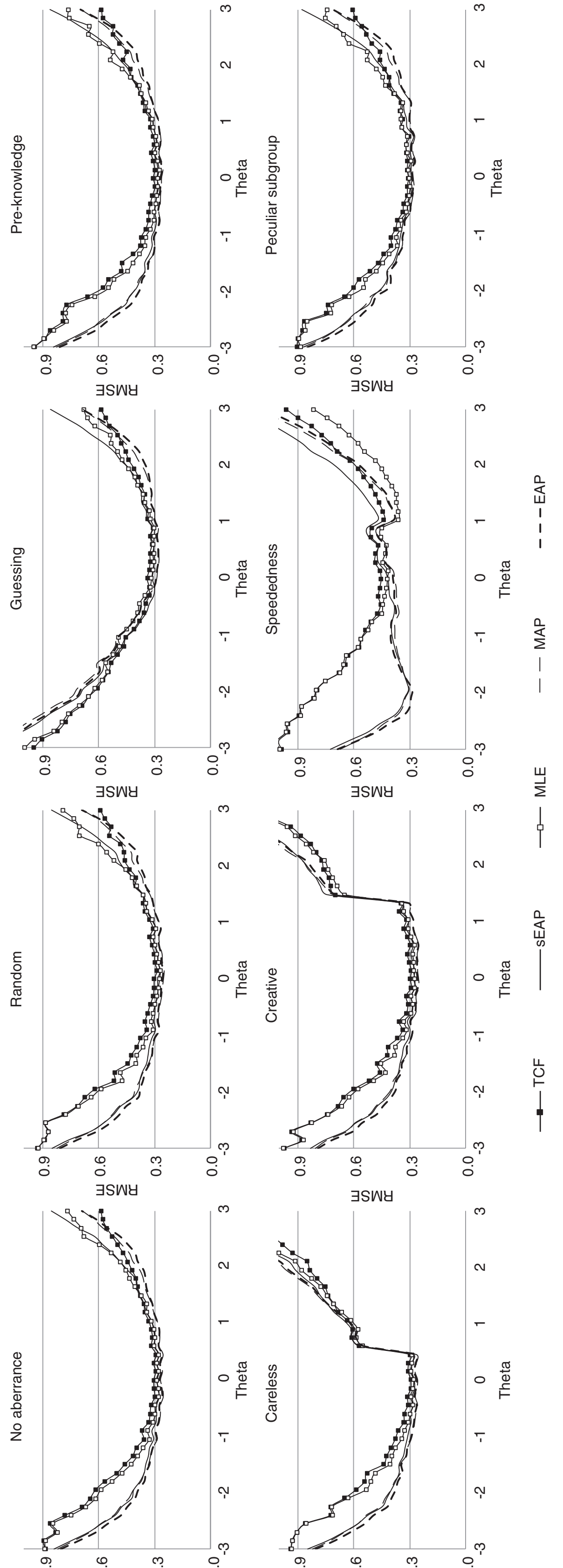

Figure 6 Conditional RMSE of each estimation method on the theta scale.

Table 3 Summary of Three Deviance Measures for Each Estimation Method Under Various Atypical Response Conditions: Theta Scale

Deviance	Method	No atypical	Random	Guessing	Preknowledge	Careless	Creative	Speededness	Peculiar subgroup
Bias	TCF	0.07	0.06	0.15	0.06	0.32	0.19	0.37	0.03
	MLE	0.04	0.05	0.17	0.04	0.33	0.19	0.33	0.05
	$_{S}$EAP	0.14	0.14	0.24	0.14	0.36	0.24	0.35	0.14
	EAP	0.10	0.10	0.23	0.10	0.36	0.23	0.30	0.12
	MAP	0.11	0.11	0.24	0.11	0.37	0.24	0.30	0.13
Error	TCF	0.35	0.36	0.36	0.36	0.36	0.35	0.36	0.37
	MLE	0.34	0.34	0.36	0.34	0.34	0.33	0.35	0.36
	$_{S}$EAP	0.29	0.29	0.31	0.29	0.30	0.29	0.29	0.31
	EAP	0.28	0.28	0.31	0.28	0.29	0.27	0.29	0.30
	MAP	0.27	0.27	0.31	0.27	0.28	0.27	0.28	0.29
RMSE	TCF	0.36	0.36	0.39	0.36	0.48	0.40	0.52	0.37
	MLE	0.35	0.35	0.39	0.34	0.47	0.38	0.49	0.36
	$_{S}$EAP	0.32	0.32	0.39	0.32	0.47	0.38	0.46	0.34
	EAP	0.30	0.30	0.39	0.30	0.46	0.36	0.42	0.32
	MAP	0.30	0.30	0.39	0.30	0.46	0.36	0.41	0.32

Notes. TCF = test characteristic function; MLE = maximum likelihood estimation; $_{S}$EAP = expected a posteriori with number-correct scoring; EAP = expected a posteriori; MAP = maximum (mode) a posteriori; RMSE = root mean square error.

Regarding the magnitude of the conditional error, all methods performed similarly across the theta region from −1.0 to +1.0 under the no aberrance, random, guessing, preknowledge, and peculiar subgroup conditions. Under those conditions, three Bayesian methods produced much less error than did non-Bayesian methods (e.g., TCF, MLE). MLE yielded the largest error in the extremes of the theta scale, whereas TCF yielded smaller error than did MLE at the upper region. Under the careless, creative, and speededness conditions, all methods performed similarly across the theta region above −1.0. However, all three Bayesian methods produced much smaller error than did their non-Bayesian counterparts at the theta region below −1.0. Under the guessing condition, TCF and MLE produced much smaller error at the lower theta region than under the other conditions. As presented in Table 3, the average weighted errors were fairly comparable across the atypical conditions. In general, the Bayesian methods produced smaller errors than did their non-Bayesian counterparts.

As a consequence of the reduction in standard error, both EAP and MAP yielded smaller RMSE than did MLE, mainly at the top and bottom theta regions, under the no aberrance, random, preknowledge, and peculiar subgroup conditions. Under the careless, creative, and speededness conditions, three Bayesian methods produced smaller RMSEs than did their non-Bayesian counterparts across only the lower region of the theta scale. There was no difference among the five methods under the guessing condition. For all atypical conditions, $_{S}$EAP produced the largest RMSEs at the upper region of the theta scale, but $_{S}$EAP performed slightly better than did TCF in the remaining region. TCF performed very similarly to MLE under the no aberrance, random, guessing, preknowledge, and subgroup conditions. TCF was more sensitive to the speededness aberrance than was MLE, however, leading to larger RMSEs in the upper theta region. MLE produced the smallest RMSEs at the top of the theta region under the speededness condition. The difference between non-Bayesian methods and their Bayesian counterparts was small under the guessing and careless conditions. As presented in Table 3, the overall RMSEs of the Bayesian methods were slightly smaller than those of their non-Bayesian counterparts. Among the three Bayesian methods, $_{S}$EAP produced slightly larger RMSEs than did EAP or MAP. All five methods produced the largest RMSEs under the careless and speededness conditions.

Reporting Score Scale

After nonlinear scaling, all three deviance measures were recalculated to compare the performance among the five proficiency estimators at the scaled score level. For each of the five proficiency estimators, we compared examinees' estimated scaled scores in a hypothetical score scale to their true scores under each of the seven atypical response conditions and the condition of no atypical responses. Figures 7–9 represent conditional bias, error, and RMSE, respectively, across the reporting score scale ranged from 130 to 170. In the figures, each plot represents each of the eight atypical conditions as in the previous figures. The bias plots display the scaled score differences (e.g., estimated minus true) associated with each

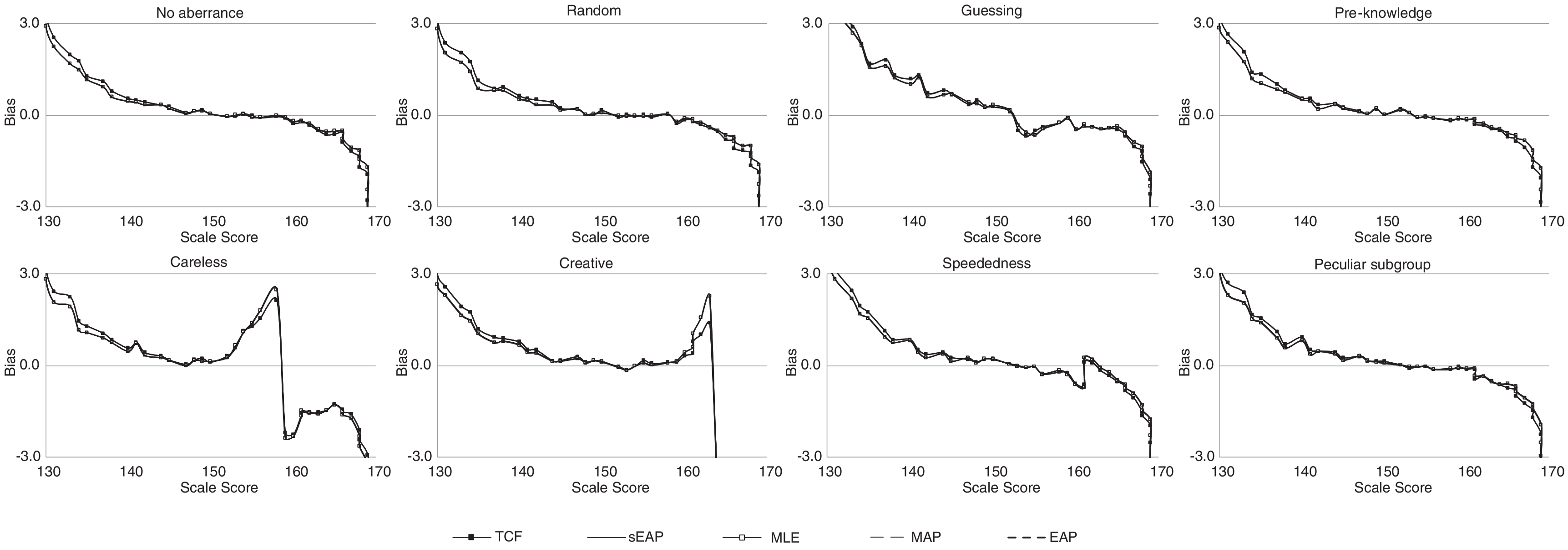

Figure 7 Conditional bias of each estimation method on the reporting score scale.

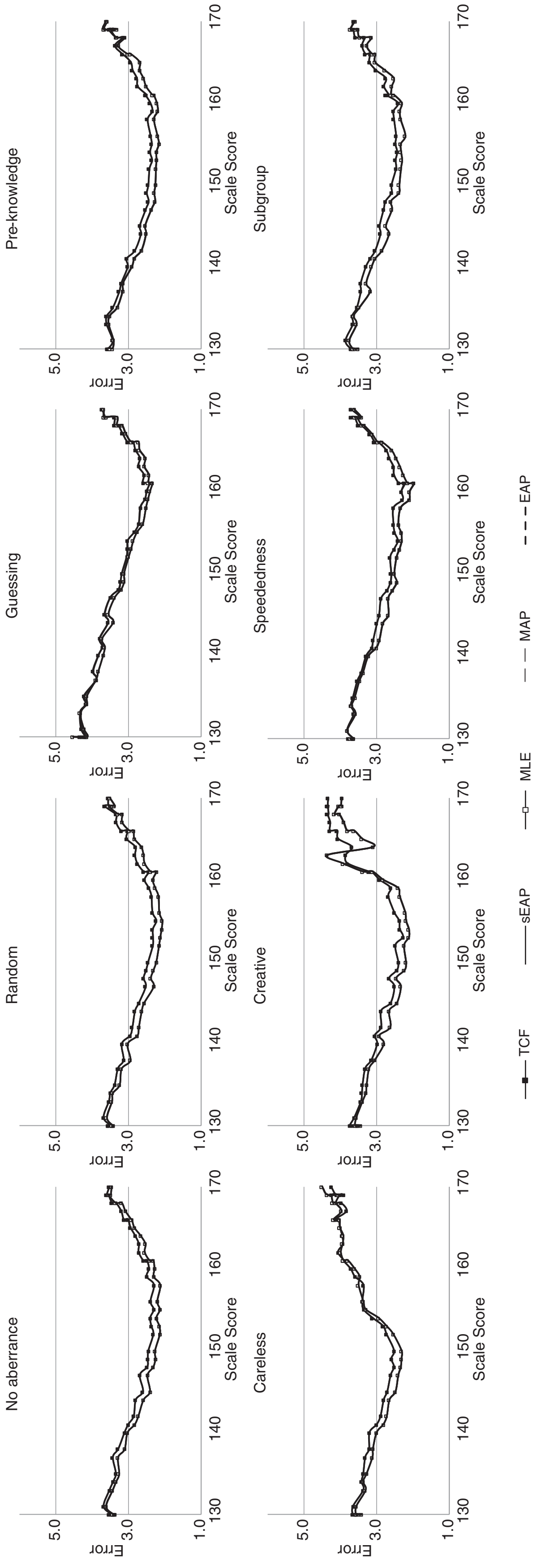

Figure 8 Conditional error of each estimation method on the reporting score scale.

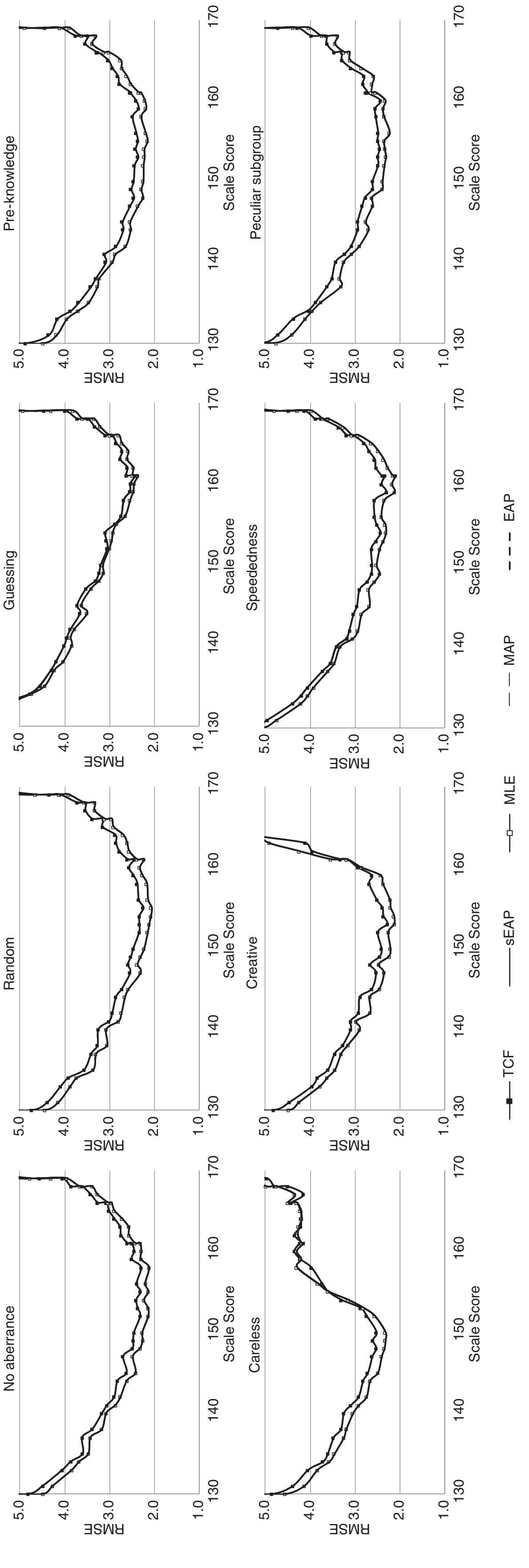

Figure 9 Conditional RMSE of each estimation method on the reporting score scale.

Table 4 Summary of Three Deviance Measures for Each Estimation Method Under Various Atypical Response Conditions: Reporting Score Scale

Deviance measure	Method	No atypical	Random	Guessing	Preknowledge	Careless	Creative	Speededness	Peculiar subgroup
Bias	TCF	0.54	0.54	0.76	0.54	1.34	1.12	0.60	0.59
	MLE	0.47	0.47	0.71	0.47	1.42	1.28	0.56	0.52
	$_S$EAP	0.54	0.54	0.76	0.54	1.34	1.12	0.60	0.59
	EAP	0.47	0.47	0.71	0.47	1.42	1.27	0.56	0.52
	MAP	0.47	0.47	0.71	0.47	1.42	1.28	0.56	0.52
Error	TCF	2.60	2.61	3.07	2.61	3.26	2.87	2.69	2.74
	MLE	2.43	2.41	2.98	2.43	3.18	2.71	2.52	2.57
	$_S$EAP	2.60	2.61	3.07	2.61	3.26	2.87	2.69	2.74
	EAP	2.43	2.41	2.98	2.43	3.17	2.71	2.52	2.57
	MAP	2.43	2.41	2.98	2.43	3.17	2.71	2.52	2.57
RMSE	TCF	2.65	2.66	3.16	2.67	3.53	3.09	2.75	2.80
	MLE	2.47	2.46	3.06	2.47	3.48	3.00	2.58	2.63
	$_S$EAP	2.66	2.66	3.16	2.67	3.53	3.09	2.75	2.80
	EAP	2.47	2.46	3.06	2.47	3.47	3.00	2.58	2.63
	MAP	2.47	2.46	3.06	2.47	3.48	3.00	2.58	2.62

Notes. TCF = test characteristic function; MLE = maximum likelihood estimation; $_S$EAP = expected a posteriori with number-correct scoring; EAP = expected a posteriori; MAP = maximum (mode) a posteriori; RMSE = root mean square error.

proficiency estimator under each of the atypical conditions. Table 4 summarizes the weighted root mean squared bias, error, and RMSE, respectively, averaged across the entire score scale from 130 to 170, derived under the eight conditions.

In Figure 7, each plot has five lines that represent the conditional bias of each of the five estimators. Only two lines are prominent in each plot, however, because the difference between Bayesian and non-Bayesian was indistinguishable in terms of conditional bias. One line indicates number-correct scoring and another indicates item-pattern scoring.[7] The difference between item-pattern scoring and number-correct scoring occurred at the two extremes of the reporting score scale, but the difference was minor. In general, all methods tended to produce positive bias in the bottom region of the scale and negative bias in the upper region of the scale across all the atypical conditions mainly due to score truncation. Both careless and creative conditions produced substantial amounts of bias, by design, in the upper score region, and as a result their averaged biases were much larger than those of the no aberrance condition. As presented in Table 4, item-pattern scoring produced slightly smaller biases than did number-correct scoring for all atypical conditions, but their differences were almost negligible. Across all the eight atypical conditions, the largest difference among the five estimators in the overall bias was 0.16 (=1.28 minus 1.12 under the creative condition).

As shown in Figures 7–9, only two patterns were clearly prominent in all plots, because the difference between Bayesian and non-Bayesian methods was indistinguishable in terms of the pattern and magnitude of conditional errors. Item-pattern scoring produced slightly smaller error than did number-correct scoring across the entire score scale for all atypical response conditions. The error patterns of the random, preknowledge, speededness, and peculiar subgroup conditions were very similar to the error pattern of the no aberrance condition. The other conditions, such as guessing, careless, and creative, produced rather different patterns compared to the no aberrance condition. Compared to the no aberrance condition, guessing produced the slightly larger error at the lower region of the scale, but both the careless and the creative conditions produced the substantially larger error at the upper region of the scale. As summarized in Table 4, item-pattern scoring produced smaller errors than did number-correct scoring, but as in the bias case, their differences were small. The overall errors were generally comparable across the eight atypical conditions. The difference between minimum (= 2.41) and maximum (= 3.26) errors was approximately one point.

As in the conditional error plots, only two patterns were clearly prominent in all plots of the conditional RMSEs for the same reason. Item-pattern scoring produced slightly smaller total error than did number-correct scoring across the entire score scale for all atypical response conditions, but the differences were minimal. The RMSE patterns of the random, preknowledge, speededness, and peculiar subgroup conditions were very similar to the RMSE pattern of the no aberrance condition. As in the error plots, guessing produced the larger RMSE at the lower region of the scale, but both careless and creative responses produced the larger RMSE at the upper region of the scale. The overall RMSEs were generally comparable across the eight response conditions. The difference between minimum (= 2.46; item-pattern scoring in the random condition) and maximum (= 3.53; number-correct scoring in the careless condition) RMSEs was slightly larger than one point.[8]

Discussion

Using real datasets (from linear tests), Kolen and Tong (2010) showed that the choice of Bayesian (prior) versus non-Bayesian (no prior) estimators was of more practical significance than the choice of number-correct (i.e., summed) versus item-pattern scoring. A recent simulation study confirmed this finding (S. Kim, Moses, & Yoo, 2015a, 2015b). In the present simulation study, the results before scaling (i.e., comparisons at the theta scale) also confirmed this trend. As shown in the no aberrance condition, the magnitude of bias and error varied depending upon the choice of IRT proficiency estimators, mainly in the top and bottom regions of the theta scale. By nature, Bayesian estimators using prior information yielded greater bias but smaller standard errors than did their non-Bayesian counterparts. For the extreme proficiency levels ($\theta < -1.5$ or $1.5 < \theta$), the decrease in standard error can compensate for the bias characteristic of Bayesian estimates. As a result, the total errors of the Bayesian methods were generally smaller than those of their non-Bayesian counterparts, but the magnitude of differences was relatively small.

IRT proficiency estimators performed differently mainly at the extremes of the theta scale as a function of atypical response types. Most estimators were robust under certain atypical conditions, where the degree of aberrance was minor either at the item level (e.g., random, preknowledge) or examinee level (e.g., peculiar subgroup). By design, about four items (i.e., two items per stage; 10% of the test) produced atypical responses under the random and preknowledge conditions, and 15% of examinees produced atypical responses under the peculiar subgroup condition. The estimation results from those atypical conditions were comparable to the result from the no aberrance condition. Again, Bayesian estimators produced smaller overall error than did non-Bayesian estimators. Both EAP and MAP performed slightly better than did $_S$EAP. Based upon the findings, we concluded that the impact of aberrance on the IRT proficiency estimators would be minimal as long as a sufficient number of high-quality items are available with which to estimate the examinees' proficiency levels. Even so, this trend will no longer be true unless the degree of aberrance is moderate, as implemented in this simulation.

Some atypical types, such as guessing, careless, creative, and speededness, resulted in larger bias (and thus larger RMSEs), compared to the no aberrance condition. Under those atypical conditions, the benefit of using prior information was almost negligible, because non-Bayesian methods performed similarly to the Bayesian methods across the theta region from −1.0 to +3.0. All estimation methods performed very similarly under each of the atypical conditions. On average, the RMSEs were the largest under the careless condition where most high-performing examinees answered easy items incorrectly due to their careless behaviors. Because many items' difficulties were lower than −0.5 in the routing module at Stage 1 (nine items out of 20) and even in the high module at Stage 2 (three items out of 20), the largest RMSEs occurring in the careless condition were not surprising. This result is compatible with the finding that the conventional 3PL-IRT model was unable to recover the (high-level) examinees' true abilities from misfit-as-incorrect responses within a test length of 45–50 items under CAT. By design (as explained in Table 2), the degree of aberrance in the creative condition was not as severe as in the careless condition, because only a few items had item difficulties lower than −1.2 at both Stage 1 and Stage 2 (either middle or high). Although the impact of both careless and creative aberrances was nontrivial to the high-performing examinees in the present study, those cases would be rare in reality. Any potential impact due to speededness must be a real concern, however, because testing programs are not completely free from this matter unless they allow unlimited time. Although we imposed a rather severe degree of speededness on the simulation, its impact was not substantially large compared to the no aberrance condition. Because difficult items appeared at the end of the test in each stage, not-reached items were more problematic to the high-performing examinees than to the low-performing examinees. By design, the examinees who were assigned to the high module at Stage 2 were not able to reach three (difficult) items out of the 40 items (one item at Stage 1 and two items at Stage 2). MLE outperformed compared to its Bayesian counterparts in the conditions where perfect scores were impossible by design (e.g., speededness, creative). MLE outperformed at the speededness (missed difficult items) condition compared to the creative (missed easy items) condition particularly for the high-performing examinees.

We used the 2PL-IRT model to estimate examinees' proficiency using known item parameters. Substantial error, caused by guessing responses, may indicate a potential concern associated with the use of the 2PL rather than the 3PL model. In this simulation, we assigned a certain correct probability to difficult items particularly for the low- and middle-level examinees. Therefore, the overall error was large across the lower region of the theta scale, and all methods performed similarly at this region. Even so, its average RMSEs were not greatly larger than the ones in the no aberrance condition. For operational purposes, at least two high-stakes international tests use the 2PL-IRT model, and the redesigned rGRE is one of them. In practice, the 2PL-IRT model will be effective when the examinees are unable to guess the correct answer without a particular ability being measured by the test. Compared to the conventional multiple-choice (MC) items that are prone to guessing, unconventional item types (e.g., text completion, sentence equivalence, or numeric entry) have been designed to prevent the examinees not only from guessing the correct answers but also from memorizing items (for test security). For example, the redesigned GRE uses various unconventional item types in the test, and often the examinees need to provide the correct answer instead of choosing an option from the multiple options. It is assumed that when the item format is cleverly designed, its correct probability caused by guessing will be very low and even close to zero.

It is expected that the contribution of the aberrant responses to measurement error on the theta scale will be replicated on the reported score scale as well. Depending on scaling conversions between the theta and reported scores, however, the relative magnitude of measurement error associated with reported scores may change in various ways across the scale. Because the reported scores ultimately matter to the test users, we further evaluated the degree of robustness of each scoring method after scaling. In general, it remained true that Bayesian methods (mainly EAP and MAP) were more effective than non-Bayesian methods, but the relatively small differences between them in the region where most examinees would likely be located became very small. After nonlinear score transformation, some differences appeared between item-pattern scoring and number-correct scoring. Across the eight atypical response conditions, item-pattern scoring performed slightly better than number-correct scoring, but in general their difference would not lead to different reported scores to the examinees (except because of rounding). Among the seven atypical types, careless and creative led to the largest errors at the score region from 155 to 170 of the reporting scale, and their conditional errors were about two or three points greater than the errors in the no aberrance condition. The sizes of total errors associated with the remaining atypical types were generally comparable to the CSEMs of the rGRE that we used as a template to create a hypothetical score scale in this simulation.

Based upon the findings derived from the comparisons at the theta scale as well as the reporting score scale, we could conclude that the impact of moderate aberrance on IRT proficiency estimates would be small as long as a sufficient number of high-quality items are available with which to estimate the examinees' proficiency levels despite some level of aberrance. If all outcomes are comparable, a simple scoring method (e.g., TCF) will be acceptable for practical. The GRE program currently uses the TCF scoring method, and this study supports the continuing use of this method in the operational setting.

In this simulation, some atypical response types (e.g., random, preknowledge, speededness, and peculiar subgroup) were not as serious as other types (e.g., guessing, careless, and creative) in terms of the extent to which the examinees' scores would change. There was no significant interaction between atypical response types and proficiency estimation methods. This means that the relative performance of the five estimation methods was generally similar across all the atypical types. There was no clear winner that is capable of precluding the impact of any aberrance from the scoring process. We concluded that possible score changes caused by the use of different proficiency estimators in a situation where certain atypical responses were present would be almost negligible.

It is unrealistic to assume that all the examinees' responses must be free from any types of aberrance. In an operational setting, many testing programs routinely carry on statistical checks to discover which operational items function differently across subgroups (e.g., item DIF) and across different time points (e.g., item drift). Problematic misfit items are not eligible for use as equaters to ensure the fairness of linking new items and also to maintain score scale stability. For example, the GRE program uses two deviance measures (the unweighted maximum difference and the weighted root mean squared error), which were designed to detect any noticeable difference between two item characteristic curves (ICCs) in terms of magnitude and pattern, to screen out misfit items from the equater set in the item calibration and linking stage. However, relatively less attention has been placed on identifying examinee characteristics that may lead to inaccurate item calibration results and furthermore score estimation results (i.e., person misfit). Many testing programs remove extremely unmotivated examinees from all the operational analyses based on a certain criterion (e.g., no response for most items).

The criterion can be further expanded to capture various types of atypical examinees as well. Screening out not only misfit items but also misfit examinees from item calibration and linking may enhance the quality and integrity of the item bank.

The focus of this study was to compare the relative performance among the five estimators under each of the eight atypical conditions. To make the present simulation manageable, we limited the study design on the basis of some information from a particular testing program and did not incorporate the following topics in the process of simulation. One limitation is related to the manner of manipulation on the examinees' responses. We manipulated the seven atypical responses that may likely occur in real testing setting. By design, certain atypical responses were rather unrealistic in severity, while others looked more realistic. Although it appeared that certain response types looked more problematic than others in terms of the estimation error, the trend will vary depending upon the degree of manipulation imposed on each of the aberrance types. Because we simulated a single level of aberrance for each of the seven atypical types, however, it is uncertain whether the same trend would appear for either higher or lower aberrance than the one imposed on this simulation. Owing to the limited study design, a definitive conclusion of the more problematic aberrance types cannot be determined from this simulation. Additional investigation is necessary to assess the impact of the percentage of atypical responses on the proficiency estimation in a solid manner.

Another limitation is related to the MST panel assembly. We used a single MST panel, which contained well-fitting items in discrimination, to generate all examinees' responses in all atypical conditions. Therefore, any investigation between IRT estimators was straightforward in this application. Even so, the use of numerous parallel MST panels in each simulation should be close to the real testing environment. It would be worthwhile to manipulate the item discrimination factor to see if using less discrimination items would confirm the current findings. A further study is needed to achieve solid conclusions on the interaction effect between MST panel type and atypical response type. Lastly, we used the 2PL-IRT model to generate item parameters and the examinees' responses to the items. Because the use of the 3PL model is common in many testing programs, comparisons between two IRT models would be interesting.

The present study supports and clarifies the findings of our previous studies (S. Kim et al., 2105a, 2015b). The most interesting feature of this study's result is that one of the IRT proficiency estimators was not particularly promising in the presence of atypical responses. The same conclusion was derived from the comparison not only at the theta scale level but also at a hypothetical reporting scale level. Based upon the findings, we recommend the use of a simpler method (e.g., non-Bayesian number-correct scoring) as a practical choice in the operational setting of the simple MST design (e.g., two-stage, overlapping modules in difficulty, etc.). In particular, no difference would be expected from the use of different proficiency estimators as long as the test includes a sufficient number of items selected from the well-maintained item bank. Despite some limitations, we think that the present findings will contribute to the MST literature. Future research geared at addressing the limitations listed above may further expand the area of MST literature.

Notes

1 Number-correct scoring is interchangeable with summed scoring.

2 Warm (1989) proposed the weighted likelihood estimator (WLE), whose estimation values are based on the mean of the likelihood function. Consequently, WLE estimates are generally slightly more central than MLE estimates, but the difference between MLE and WLE would be small.

3 The range of values for the estimated proficiency level is usually restricted so that the maximum likelihood method provides finite estimates. We set an upper bound (+5.0) for the only-correct score as well as a lower bound (−5.0) for the only-incorrect score.

4 The *defined population intervals* method can be used to implement a policy that specifies the relative proportions of examinees in the population expected to follow each of the available paths through the panel.

5 We simulated the examinees' theta distribution to be uniform to obtain stable estimates over the entire theta scale. To compute the summary statistics, however, we applied the normal distribution weights so as to reflect a more realistic distribution of examinees' thetas.

6 For simplicity, we used the standard normal distribution as the population distribution.

7 The scaling procedure used in this study was based on a similar form of equipercentile function applied to a single set of population theta probabilities. As a result, scaling results mainly differed by how many estimated thetas an estimation method produces. For example, number-correct scoring produces 41 estimated thetas (we used a 40-item test), whereas item-pattern scoring produces numerous estimated thetas. All three deviance measures mainly reflect whether the scaled estimates are based on number-correct scoring or item-pattern scoring.

8 The reporting scale used in this study is hypothetical but its properties are very similar to the rGRE scale's properties. As appeared in the *GRE Guide to the Use of Scores* (2014 – 2015), the CSEMs at the scaled scores of 140 – 160 ranged from 2 to 3 for both verbal and quantitative measures. In light of the CSEM size, the magnitude of the difference can be considered small.

References

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York, NY: Marcel Dekker.

Baker, F. B., & Kim, S. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431 – 444.

Chang, H., & Ying, Z. (2009). To weight or not to weight? Balancing the influence of initial items in adaptive testing. *Psychometrika, 73,* 441 – 450.

Educational Testing Service. (2011). *GRE information and registration bulletin*. Princeton, NJ: Author. Retrieved from http://www.ets.org/s/gre/pdf/gre_info_reg_bulletin.pdf

Educational Testing Service. (2014). *GRE Guide to the use of scores (2014 – 2015)*. Princeton, NJ: Author. Retrieved from http://www.ets.org/gre/institutions.

Guyer, R. D., & Weiss, D. J. (2009). Effect of early misfit in computerized adaptive testing on the recovery of theta. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved from www.psych.umn.edu/psylabs/CAT Central/

Haberman, S. J. (2006). *An elementary test of the normal 2PL model against the normal 3PL alternative* (Research Report No. RR-06-14). Princeton, NJ: Educational Testing Service. 10.1002/j.2333-8504.2006.tb02020.x

Ho, T.-H., & Li, F. (2013, April). *To Bayes or not to Bayes: Reducing the impact of response anomalies in adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), San Francisco, CA.

Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education, 19*(3), 203 – 220.

Karabatsos, G. (2003). Comparing the atypical response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16,* 277 – 298.

Kim, H., & Plake, B. S. (1993, April). *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Kim, S., & Moses, T. (2014). *An investigation of the impact of misrouting under two-stage multistage testing: A simulation study* (Research Report No. RR-14-01). Princeton, NJ: Educational Testing Service. 10.1002/ets2.12000

Kim, S., Moses, T., & Yoo, H. (2015a). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement, 52*(1), 70 – 79.

Kim, S., Moses, T, & Yoo, H. (2015b). *Effectiveness of item response theory (IRT) proficiency estimation methods under adaptive multistage testing* (Research Report No. RR-15-11). Princeton NJ: Educational Testing Service. 10.1002/ets2.12057

Kim, S., Robin, F., & Liu, J. (2012). *An empirical investigation of the impact of misrouting under two-stage multi-stage testing of rGRE®* (Statistical Report No. SR-2012-020). Princeton, NJ: Educational Testing Service.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practice* (2nd ed.). Berlin, Germany: Springer Science + Business Media.

Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement Issues and Practice, 29,* 8 – 14.

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet-assembly design for adaptive multistage tests. *Applied Measurement in Education, 19,* 189 – 202.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35,* 229 – 249.

Luecht, R. M., Nungester, R. J., & Hadidi, A. (1996, April). *Heuristic-based CAT: Balancing item information, content and exposure.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Luecht, R. M., & Sireci, S. G. (2011). *A review of models for computer-based testing* (College Board Research Report No. 2011 – 12). New York, NY: College Board.

Magis, D., Béland, S., & Raîche, G. (2011). A test-length correction to the estimation of extreme proficiency levels. *Applied Psychological Measurement, 35,* 91 – 109.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18,* 311 – 314.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9,* 3 – 8.

Meijer, R. R. (2005). *Using patterns of summed scores in paper-and pencil tests and CAT to detect misfitting item score patterns* (Computerized Testing Report 02–04). Newtown, PA: Law School Admission Council.

Robin, F. (2014). *Estimation bias and scoring bias.* Unpublished memorandum.

Rulison, K. L., & Loken, E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement*, *33,* 83–101.

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66,* 191–207.

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20,* 227–253.

Tong, Y., & Kolen, M. J. (2010, April). *IRT proficiency estimators and their impact*. Paper presented at the annual meeting of the National Council in Measurement in Education, Denver, CO.

Wang, X., Fluegge, L., & Luecht, R. (2012, April). *A large-scale comparative study of the accuracy and efficiency of ca-MST.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika, 54,* 427–450.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed. pp. 111–153). Westport, CT: American Council on Education and Praeger.

Yen, Y., Ho, R. Laio, W., Chen, L., & Kuo, C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. *Applied Psychological Measurement, 36,* 75–87.

Suggested citation:

Kim, S., & Moses, T. (2016). *Investigating robustness of item response theory proficiency estimators to atypical response behaviors under two-stage multistage testing* (GRE Board Research Report No. 16-03, ETS Research Report No. RR-16-22). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12111

Action Editor: Brent Bridgeman

Reviewers:

This report was reviewed by the GRE Technical Advisory Committee and the Research Committee and Diversity, Equity and Inclusion Committee of the GRE Board.