Designing CAT MOCCA: Guiding Principles and Simulation Research

MOCCA Technical Report MTR-2021-1

Mark L. Davison
David J. Weiss
Ozge Ersan
Joseph N. DeWeese

University of Minnesota

Gina Biancarosa
Patrick C. Kennedy

University of Oregon

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A140185 and R305A190393 to the University of Oregon. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Contents

List of Figures

List of Tables

Executive Summary

MOCCA is an online assessment of inferential reading comprehension for students in 3rd through 6th grades. It can be used to identify good readers and, for struggling readers, identify those who overly rely on either a Paraphrasing process or an Elaborating process when their comprehension is incorrect. Here a propensity to over-rely on Paraphrasing or Elaborating is called a Process Propensity. MOCCA is diagnostic of reading comprehension difficulties, and it can be used as an outcome measure, as a formative assessment, or as a progress monitoring tool. To improve MOCCA diagnostic capabilities and minimize testing time, the test has been converted to a computerized adaptive testing (CAT) format. This report describes a series of monte-carlo simulation studies conducted to guide decisions about the design of MOCCA CAT.

To improve MOCCA, we have adopted a sequential, variable-length CAT testing process. In Phase 1, students are administered MOCCA items to locate the student along the Reading Comprehension dimension. Based on the student's Reading Comprehension score, we decide if the student is a reader whose instruction could benefit from knowledge of the student's Process Propensity. Such readers are here defined as students with a score below the mean of our calibration sample (Davison et al., 2019). For above average readers, testing ends with Phase 1. For below average readers, testing proceeds to Phase 2, in which the student is administered additional items for purposes of determining whether the student has a Process Propensity and, if so, if it is a propensity toward the paraphrasing process or the elaborating process.

Phase 1

In the simulation for the design of Phase 1, we studied three independent variables: stopping rule, upper limit of test length, and number of item options. There were four dependent variables: bias, root mean square error (RMSE), root mean square standard error of measurement (RMS-SEM), and average test length (ATL) of the variable-length CAT. The first independent variable was a stopping rule with two levels: stop when the student's estimated standard error of measurement falls below 0.30 vs. when the student's estimated standard error falls below 0.35. Within the constraints of our item bank and feasible test lengths, the stricter stopping rule, SEM = 0.30, led to increased test length with little improvement in average bias, RMSE, or RMS-SEM. Therefore, we have adopted a stopping rule of SEM $\leq$ 0.35.

Our second independent variable was the maximum number of items for Phase 1, with three levels: 25, 30, and 40. Once most students had taken 25 items, additional items yielded diminishing returns in terms of improved bias, RMSE, and RMS-SEM. Therefore we have adopted an upper limit of 25 items in Phase 1.

Our third independent variable was the number of item options: 3 vs. 5. Items with five options and a reduced probability of guessing resulted in better performance of the CAT, particularly with respect to the RMS-SEM and test length. Therefore, as we develop new items, we are including items with five options.

Phase 2

For Phase 2, we investigated five independent variables: (1) administration of Phase 2 (administration to all students vs. administration to below average readers only), (2) two kinds of information statistics for selecting the next item in the CAT (Fisher information vs. weighted Fisher information), (3) three classification rules [confidence interval rule (CI), sequential probability ratio test (SPRT), and generalized likelihood ratio rule (GLR)], (4) width of indifference region for the GLR ([−0.25. 0.25], [−0.50, 0.50], and [−1.00, 1.00]), and (5) test length with two levels (25 items Phase 1, 15 items Phase 2 vs. 25 items Phase 1, 40 overall). There were two primary dependent variables: classification accuracy and test length.

Administering Phase 2 only to below average readers led to a substantial reduction in test length for some simulees. It also improved classification accuracy. Therefore, the CAT will end after Phase I for those whose estimated Reading Comprehension is above the mean; only students whose Reading Comprehension scores are below the mean will proceed to Phase 2.

As compared to Fisher information, our weighted Fisher information led to better classification accuracy and shorter test lengths, but primarily for simulees with true Reading Comprehension scores above the mean, few of whom will enter Phase 2. Because it led to some improvements, we decided to use weighted Fisher information in the CAT. In comparisons of classification rules, the GLR and SPRT rules had better classification rates and shorter test lengths than did the CI rule. Because the GLR displayed slightly better performance and has a stronger theoretical rationale, we decided to implement the GLR in CAT MOCCA.

The width of the GLR indifference region did not yield differences that favored one width over the others across the dependent variables. The indifference region [−0.50, 0.50] offered a good compromise, and we are adopting it as the indifference region for MOCCA CAT.

Finally, for the test length options (25/40 vs. 25+15), the 25/40 option resulted in slightly longer tests, but higher classification accuracy. In our view, the higher accuracy justifies the additional length, so we have adopted the 25/40 rule: a limit of 25 items for Phase 1 and a total test limit of 40 items.

Our final CAT is a variable-length CAT with two phases. In Phase 1, students take MOCCA items to measure their Reading Comprehension. At each step of Phase 1, the item administered is the one with the highest Fisher information for the Reading Comprehension dimension from among the items not yet administered. Testing proceeds until the student's estimated standard error of measurement falls below 0.35 or the number of administered items reaches 25, whichever comes first. If the student's Phase 1 Reading Comprehension score is above the mean for our calibration sample, testing will end with Phase 1. If their Reading Comprehension score is below the mean, the student will proceed to Phase 2. At each step of Phase 2, the student's estimated Process Propensity score will be based on incorrect responses from Phase 1 and Phase 2 combined. At each step of Phase 2, the administered item is the one with the highest weighted Fisher information for the Process Propensity dimension from among those not yet administered.

Testing stops when the student's comprehension process is classified as either Paraphrasing or Elaborating using the generalized likelihood ratio rule and an indifference region of [−0.50, 0.50] or the student has taken a total of 40 items in Phase 1 and Phase 2 combined. The student's score report will include a Reading Comprehension score and a Process Propensity classification, if the student has been classified in Phase 2, but they will not receive a numeric score for the Process Propensity dimension. The Process Propensity classification is designed to be a qualitative description useful in the design of future instruction for that student.

Designing CAT MOCCA: Guiding Principles and Simulation Research

The primary purpose of this technical report is to describe the extensive simulation research used to guide the development of the computerized adaptive test (CAT) format of MOCCA, the Multiple-Choice Online Causal Coherence Assessment of inferential reading comprehension. A second purpose is to describe the reading assessment and measurement principles underlying the CAT MOCCA assessment. MOCCA was originally developed as a linear, paper-pencil assessment, evolved into a computer-administered assessment for 3rd through 5th graders, and now has further evolved into a CAT assessment of inferential reading comprehension for 3rd through 6th graders. Before describing the simulation studies guiding the development of the CAT MOCCA assessment, this report provides context for that research in the form of background on MOCCA, the theoretical conceptualization of the reading process that has guided its development, and descriptions of the assessment, its intended uses, and the improvements in MOCCA we hoped to achieve by converting the assessment to a CAT format.

Theoretical and Measurement Foundations of MOCCA

MOCCA was originally designed to measure overall inferential reading comprehension in grades 3 through 5 and to provide diagnostic information as to why some students struggle with comprehension. Struggling readers differ in the reasons why they struggle. They might struggle with decoding or with reading words accurately and fluently. They might have limited English vocabulary or background knowledge. However, some students struggle with comprehension itself, despite having adequate decoding, word reading, vocabulary, and background knowledge.

Readers who struggle with comprehension itself struggle to make inferences that help maintain a coherent idea of what a text is about. These struggling readers are usually trying to make sense of what they read but are relying on strategies that are not effective. In practice, they tend to rely on one of two strategies: paraphrasing or elaborating. The paraphrasing strategy limits the understanding to what is explicitly stated in the text, but comprehension requires an inference beyond the explicit material in the text. The elaborating strategy permits inferences beyond the text. It includes elaborative inferences, personal associations, evaluations, self-explanations, and references to background information, but it does not lead to the inference that results in comprehension.

Depending on the context, most readers will use both elaborating and paraphrasing strategies, but many readers tend to rely on one of the two strategies more than the other. While both are good strategies, neither alone will necessarily result in good comprehension. Research suggests that students who predominantly rely on paraphrasing require somewhat different instruction than do students who predominantly rely on elaborating (Liu, Kennedy, Seipel, Carlson, Biancarosa, & Davison, 2019; McMaster, Espin, & van den Broek, 2014; McMaster, van den Broek, Espin, White, Kendou, Rapp, Bohn-Getter, & Carlson, 2012; Rapp, van den Broek, McMaster, Kendeou, & Espin, 2007). MOCCA is designed to be diagnostic in that it helps distinguish between struggling comprehenders who rely predominantly on the paraphrasing strategy versus those who rely predominantly on the elaborating strategy.

In terms of psychometric theory, there are several ways to conceptualize the distinction between the two types of struggling readers. In our work, we have conceptualized the distinction

in terms of a continuous, bipolar dimension that we call the Process Propensity (PP) dimension. At the negative end of the dimension are readers who rely solely on the elaborating strategy when they struggle with comprehension. At the positive end are readers who rely solely on a paraphrasing strategy when they struggle with comprehension. In the middle, around the zero point of the dimension, are readers who use the two strategies equally. Thus, we view elaborating and paraphrasing strategies not as discrete categories, but as ends of a bipolar, continuous dimension that reflects the student's propensity to use the paraphrasing strategy, rather than the elaborating strategy, or vice-versa, when struggling with comprehension.

The Nature of MOCCA

In the non-adaptive, computer-administered version for 3rd – 5th graders, MOCCA consists of 40 multiple-choice items. Each item consists of a short story in which a sentence is missing. From three alternative responses, the student must choose the sentence that best completes the story. Whereas most multiple-choice items include two types of response alternatives, correct and incorrect, each MOCCA item includes three types of response alternatives: correct, paraphrase, and elaboration.

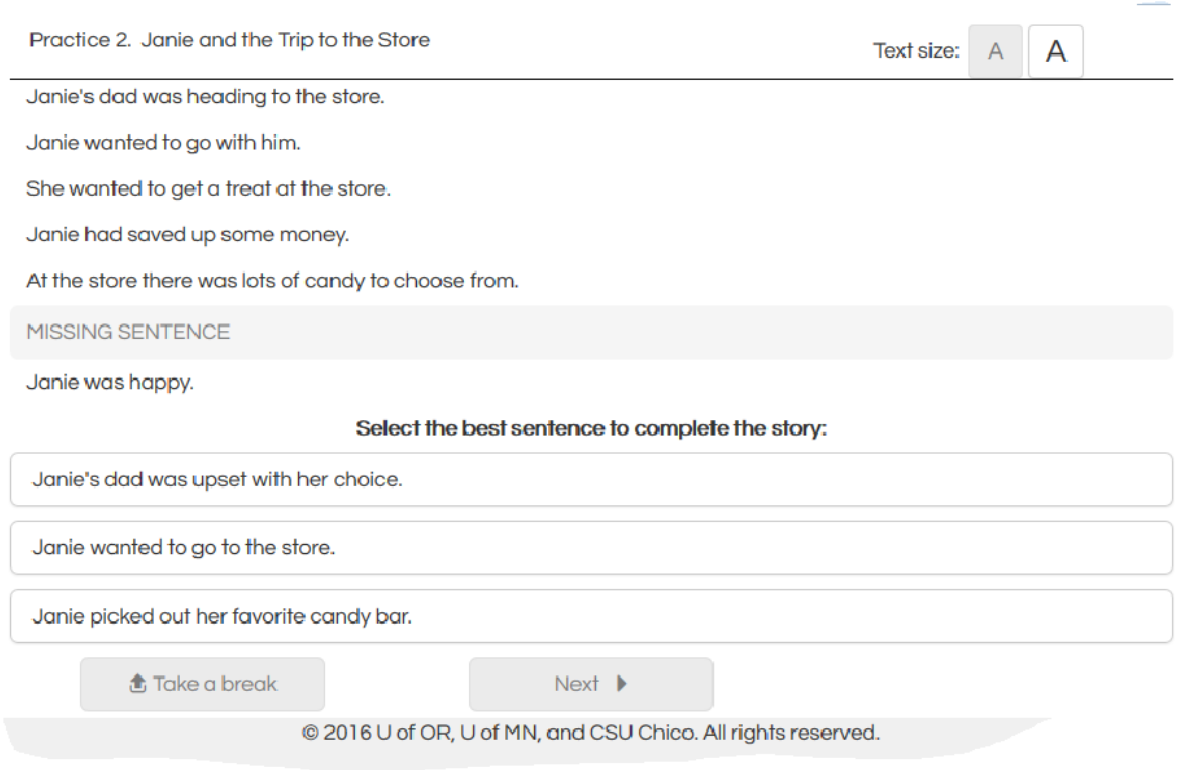
Figure 1 shows a sample MOCCA item entitled "Janie and the Trip to the Store." Note how the sixth sentence is missing, and that there are three sentences at the bottom, representing three possible responses for the missing sentence. The first alternative "Janie's Dad was upset with her choice." is the elaboration response. It states information not explicitly stated in the passage, and therefore involves an inference, but it does not complete the story, because it is inconsistent with the last sentence. The second sentence "Janie wanted to go to the store." is the paraphrase response that merely reiterates information explicitly stated earlier in the story. The third alternative is the correct response: "Janie picked out her favorite candy bar." It is an inference, in that it states information not stated earlier, and it completes the story in that it explains why Janie is happy in the last sentence, and it states whether she accomplished her goal of getting a treat.

In addition to containing three types of alternatives for each item, MOCCA items differ from those usually seen in reading comprehension tests in one other important respect. Many reading tests contain passages with several items related to each passage. Since the several items for a single passage all refer to the same passage, they form testlets that might violate the local independence assumption of item response theory (IRT). In MOCCA, there is only one item for each story, so the structure of the item does not impose violations of the local independence assumption. The independence of items means that MOCCA items satisfy the IRT assumptions of independence and makes MOCCA highly suitable for a computerized adaptive format, particularly as compared to other reading tests.

All stories were reviewed for cultural and developmental appropriateness, among other things, by an external panel of six teachers who worked with Grade 3–5 students, including a special education teacher and a Title 1 specialist from a Spanish-English dual-language school. Items flagged by the teachers were reviewed and revised or dropped. Stories were then selected to balance forms within grade by readability as measured by Flesch-Kincaid Grade Level (Kincaid, Fishburne, Rogers, & Chissom, 1975); within a grade, the average Flesch-Kincaid level of the stories is held constant across fixed-item forms, but increases systematically over grades. A range of other story features were also balanced across forms, such as the gender of the main character, the explicitness of the goal, and whether the end of the story satisfied the main goal or not. All items were reviewed for differential item functioning by gender and race/ethnicity using the

Mantel-Haenszel statistic (Holland & Thayer, 1986). Due to subsample size limitations, the DIF analysis by race/ethnicity could only be conducted for the White Hispanic and White non-Hispanic ethnicity groups. See Davison, Biancarosa, Seipel, Carlson, Liu, and Kennedy (2019) for more detail on the construction and validation of MOCCA.

Figure 1. Sample MOCCA Item



Intended Uses

General Uses. MOCCA CAT is designed to be a diagnostic assessment for struggling readers in 3rd – 6th grades. Use beyond these grades is not recommended because validity evidence has not been collected beyond these grades. MOCCA CAT is diagnostic in that it can provide information not only about inferential reading achievement level, but also diagnostic information about the comprehension processes used by struggling readers. MOCCA CAT is designed to provide information that can be used to assess general reading comprehension ability, identify the comprehension processes used by struggling comprehenders (i.e., paraphrasing, elaborating strategies), determine comprehension efficiency (i.e., fast or slow), inform instruction, monitor progress, and make response-to-intervention decisions.

Specific Uses. MOCCA CAT has a large item pool. It can be administered up to five times without a student seeing any item twice. This makes it useful for benchmarking progress over time, either progress within a grade or progress from grade-to-grade in 3rd – 6th grades, although it has not been thoroughly validated for benchmarking progress.

The earlier, non-adaptive computer-administered versions of MOCCA have concurrent and construct validity evidence to indicate that MOCCA provides information similar to that provided

by other more-traditional reading tests (Biancarosa, Kennedy, Carlson, Yoon, Seipel, Liu, & Davison, 2019; Davison et al., 2019). Additionally, the non-adaptive computer-administered versions of MOCCA have been validated as a cognitive diagnostic tool for identifying at-risk students (Biancarosa et al., 2019; Davison et al., 2019; Liu et al., 2019; Su & Davison, 2019). Davison et al. (2019) report evidence on the concurrent validity of the non-adaptive computer editions for purposes of screening and benchmarking. In the future, this validity research will be extended to MOCCA CAT.

Inappropriate Uses. As with any benchmarking or diagnostic reading comprehension measure, MOCCA CAT is best used in combination with other assessments to provide a more complete description of a child's reading achievement. MOCCA does not provide diagnostic information about decoding or other, more basic, early component reading skills.

Why an Adaptive Format for MOCCA?

Prior work with MOCCA identified two possible areas for improvement that could be addressed using a CAT format: diagnostic accuracy and test administration efficiency. First, for MOCCA to reliably and accurately separate poor comprehenders into diagnostic groups, standard errors for the Process Propensity dimension must be sufficiently small, which means readers must get several items incorrect. For example, if a reader makes only three errors on the 40-item test, we have only three data points to assess the reader's propensity to paraphrase rather than elaborate when choosing an incorrect response. Second, the conventional fixed-form versions of MOCCA take an average of 35 minutes to complete but can take some students 50 minutes or more. Given prevailing attitudes regarding "over-testing" students, a 50-minute administration, and possibly even a 30-minute administration, is inefficient and undesirable. These two issues led us to develop a CAT format for MOCCA to reduce testing time and increase the classification accuracy of poor comprehenders by process propensity type.

Our approach to MOCCA CAT is a sequential approach. The purpose of Phase 1 is to estimate the student's location along the Reading Comprehension (RC) dimension. Thus, in Phase 1, administered items are chosen to optimize information regarding Dimension RC. Students proceed to Phase 2 only if their Dimension RC score in Phase 1 is below the mean of our calibration sample (Davison et al., 2019). This cut-off was chosen for three reasons. First, readers at or above the mean have missed only a few items, too few to be classified with what we consider sufficient accuracy. Second, readers above the mean are unlikely to need supplemental instruction, and therefore, for above average readers, a classification as paraphrasing or elaborating is unlikely to be used in the design of supplemental instruction. Third, our goal is to provide a classification for students that are considered by their teacher to be a struggling reader and in need of supplemental instruction. By providing a classification for students below the mean on Dimension 1, the test should provide a classification for almost all students who demonstrate a propensity for either paraphrasing or elaborating and who are considered in need of supplemental instruction (a struggling reader) by their teacher, although the test will also provide a classification for some readers who score below the mean but are not considered by their teacher to be either struggling or in need of supplemental instruction. We are not suggesting that all students below the mean are struggling, but we do think that almost all students in need of supplemental intervention will score below the mean on MOCCA.

In Phase 2, additional items are chosen to be optimal for the assessment of the Process Propensity (PP) dimension. Because assignment to an intervention is a classification decision, we

used classification CAT to provide a classification of students' comprehension process as (1) predominantly a paraphrasing process, (2) predominantly an elaborating process, or (3) inconclusive. These classifications are explained in more detail below. Thus, in Phase 2, the student takes additional items to improve classification accuracy. Using the items the student answered incorrectly in Phase 1, an initial estimate of the student's location along Dimension PP is computed as the initial estimate for Phase 2. Then additional items are selected and administered to improve the student's process propensity classification.

Three primary goals motivated our efforts to convert MOCCA to a CAT format. First, we aimed to minimize the number of items needed in Phase 1 to estimate students' locations along the Reading Comprehension dimension. CAT is a machine learning procedure in which the computer gradually learns how to select optimal items for a given student. Selecting optimal items optimizes testing by (1) reducing the number of items needed to reach a fixed level of precision (standard error of measurement) or (2) increasing the precision of measurement for a fixed test length (Weiss, 1982). Our goal was to minimize the number of items administered in Phase 1 to make it possible to administer an adequate number of items in Phase 2 while keeping overall test length at or below 40 items for all students.

Second, we wanted to control and equalize the standard error of measurement in Phase 1 to the extent possible. To achieve this, we adopted a variable-length CAT approach for Phase 1. Using this approach, each student is tested until their standard error of measurement is equal to or less than a preset cut-off value, which nearly equalizes the standard error of measurement for all students, as all students will have standard errors at or below the pre-set level, unless their score does not reach the pre-set level of precision within the maximum number of items allowed, which most frequently results from limitations inherent in the CAT item bank.

Third, we wanted to improve classification of readers' comprehension processes into either the elaborating, paraphrasing, or inconclusive categories. Reducing the number of items needed to accurately place students along the reading comprehension dimension would mean that, without increasing test length beyond 40, we could administer some items solely for the purpose of diagnostic classification. These would be somewhat more difficult items, because the student's response provides information about error propensity only if the student incorrectly answers the item.

Methods

The purpose of these simulation studies was to design a CAT procedure that would yield accurate estimates of a student's reading comprehension ability and process propensity classification. Therefore, the sole focus was on estimation of person parameters and person classifications. The CAT is delivered sequentially, with Phase 1 devoted to estimation of reading comprehension ability (i.e., Dimension RC), and Phase 2 devoted to process propensity classification (i.e., Dimension PP). Thus, this section is divided into two parts, one describing the methods for implementing measurement CAT for Dimension RC in Phase 1 and the second describing the methods for implementing classification CAT on Dimension PP in Phase 2.

Phase 1

Model. Our model for Dimension RC and Dimension PP is a two-dimensional, linear tree model (De Boeck, Chen, & Davison, 2017; De Boeck & Partchev, 2012; Partchev & De Boeck,

2012). Because such models are not widely known and, to our knowledge, have never been applied to correct and multiple incorrect response type data, our linear tree model is explained in some detail.

In tree models, an item response generates a small vector of response variables, not a single response variable. In our case, there are two response variables for each item *j*, $\boldsymbol{X}_{ij} = (X_{1ij}, X_{2ij})$. The first of these is the focus in Phase 1, and the second is the focus of Phase 2 and is discussed below (e.g. Equation 8). The first is the familiar correct and incorrect variable X_{1ij}, response variable 1 for person *i* (*i* = 1, …, *I*) and item *j* (*j* = 1, …, *J*):

$$\begin{aligned} X_{1ij} &= 1 \quad \text{if the response of person } i \text{ to item } j \text{ is correct} \\ &= 0 \quad \text{if the response of person } i \text{ to item } j \text{ is incorrect} \end{aligned} \tag{1}$$

For the first response variable of Equation 1, we assumed a three-parameter logistic (3PL) model in which $\theta_{i.RC}$ is the person parameter locating person *i* along Dimension RC, but we constrained lower asymptote parameters to be equal across items. Let $\alpha_{RC.j}$, $\beta_{RC.j}$, and c_j be the discrimination, difficulty, and guessing parameters, respectively, for Dimension RC and item *j*. Then our model for Dimension RC is the familiar 3PL model:

$$\pi_{1ij}(X_{1ij} = 1) = c_j + (1 - c_j)\left[\frac{\exp\left[\alpha_{RC.j}(\theta_{i.RC} - \beta_{RC.j})\right]}{1 + \exp\left[\alpha_{RC.j}(\theta_{i.RC} - \beta_{RC.j})\right]}\right] \tag{2}$$

Assuming that the responses to any two items, *j* and *j'* are independent, conditional on θ_{RC}, the likelihood of the response vector for person *i*, $\boldsymbol{x}_{1i} = (X_{1i1}, X_{1i2}, \ldots, X_{1iJ})$ is

$$L_{i.RC} = \prod_{j=1}^{j=J} (\pi_{1ij})^{X_{1ij}} (1 - \pi_{1ij})^{1 - X_{1ij}} \tag{3}$$

For this model, the maximum likelihood estimate of θ_{RC} is the value that maximizes this likelihood function. Given that maximum likelihood estimates can be undefined for some response vectors, we estimated θ_{RC} using weighted maximum likelihood (Warm, 1989), using unidimensional item response theory software (Xcalibre; Guyer & Thompson, 2014)

For the item parameters, we used the real test item parameters for 360 items developed for the computerized, but non-adaptive, edition of MOCCA. These parameters, estimated by Xcalibre, are shown in Appendix Table A.1 along with model fit statistics (Table A.2), bank information function plots (Figure 3), and conditional standard error plots for the bank (Figure 3). For the RC Dimension, the mean and standard deviation of the discrimination ($\alpha_{RC.j}$) parameters are 1.899 and 0.394, and for the difficulty ($\beta_{RC.j}$) parameters they are –0.217 and 0.440. Guessing parameters (c_j) were fixed at 0.24 for all items. For the PP Dimension, the mean and standard deviation of the discrimination parameters are 1.171 and 0.173, while those for the difficulty parameters are –0.350 and 0.558.

The items in the real item bank, whose parameters provided the basis for this study, were calibrated using samples described in Davison et al. (2019). Davison et al. (2019) also describes the item development process, the anchor item selection process, and the test administration procedures. In the calibration, there were three forms at each grade for a total of nine forms, each with

40 items. The total sample size was 5,866 with more than 400 students taking each non-anchor item and over 1,000 taking each anchor item. Because of problems in estimating parameters for Dimension PP in the original study, the parameters were re-estimated using Xcalibre (Guyer & Thompson, 2014) with boundaries on parameters and informative priors.

A 3PL model (Equation 2) with all guessing parameters constrained equal was specified for the RC responses. A 2PL model ($c_j = 0$ in Equation 2) was specified for the PP response variable. For both dimensions, the prior distribution for the difficulty was specified with mean 0 and variance 1.00, the prior distribution for the discrimination was specified with mean 2.00 and standard deviation 0.25. For the RC Dimension, the guessing parameters were specified as 0.24 based on pilot analyses in which most items had lower asymptotes near 0.24. For both analyses, maximum and minimum boundaries for the discrimination parameter were set at 6.00 and 0.05, and at 4.00 and –4.00 for the difficulty parameters. The calibration was a concurrent calibration across grades and forms with nine anchor items.

For the person parameters in the simulation studies, θ was assumed to take on 15 values from –2.8 to +2.8 in increments of 0.4 along Dimension RC. We specified 500 simulated examinees (simulees) at each θ value for a total of 7,500 simulees. Although this discrete and uniform distribution of θ is not likely to occur in practice, we chose to simulate θ values in this fashion so we could study how the dependent variables varied conditional on θ. Given that θ_{RC} is estimated independently for each simulee, the estimate for a given simulee should not be affected by the distribution of θ_{RC} for other simulees.

Independent Variables. CAT requires two decision rules—an item selection rule to decide which item to administer next, and a stopping rule to decide when to stop testing. Most of our independent variables involved comparisons of different item selection or stopping rules.

During this investigation, one of the independent variables became moot. Prior investigation revealed that the item bank derived earlier in the MOCCA project was less than ideal for CAT. Along both dimensions, item difficulties were heavily concentrated around the dimension mean of zero. In other words, there were too many items of moderate difficulty, and not enough items at extreme difficulties on either dimension. Concurrent with this simulation research, item developers were attempting to write new items with more extreme difficulties so that the item bank would contain items with difficulties that better spanned the full range of both dimensions. Early in the simulation research, the study included a "real" item bank condition, in which the item difficulties had the same limited range as in our existing items, and an "ideal" item bank in which the item difficulties spanned the full range of difficulties. Both "real" and "ideal" banks were studied so that we would have data to design the CAT administration regardless of whether the new item writing efforts were successful in creating a bank whose difficulties better spanned the full range of θ.

As pilot data on the new items were collected, it became clear that the new items spanned only a slightly wider range of difficulties. While we do report some data on the gains in measurement precision that could be achieved with a more ideal item bank, for the sake of parsimony, the major independent variables were studied only in the context of the "real" item bank. This is because the actual item bank available to the CAT will more nearly resemble the "real" item bank. We hypothesize that the tight constraints the MOCCA item development process places on many item features (e.g., Flesch Kincaid level, number of sentences, sentence length) might limit the range of item difficulties. After dropping the "ideal" vs. "real" item bank factor, there were three remaining independent variables related to Phase 1: (1) the standard error of measurement stopping rule, (2) the upper bound on number of items in Phase 1, and (3) the number of response options.

Stopping Rule: Estimated Standard Error of Measurement (SEM). First, we varied the estimated SEM stopping rule, a factor with two levels: 0.35 and 0.30. In the first condition, testing in Phase 1 ended when the estimated SEM reached 0.35 or below or when the number of items reached 25, whichever came first. In the second level of the factor, testing in Phase 1 ended when the estimated SEM reached 0.30 or below or when the number of items reached 25, whichever came first.

Upper Bound on Number of Items. Second, we varied the maximum number of items that a person could take in Phase 1. Based on the examination of SEM stopping rules, we decided to use a stopping rule of 0.35. Three levels of this factor were studied: (1) 40 items, (2) 30 items, and (3) 25 items. In all three conditions, testing in Phase I stopped when the SEM reached 0.35 or below or when the number of items reached the upper limit (either 40, 30, or 25 items).

Number of Response Options. Finally, we varied the number of response options. Historically, MOCCA items have had three response options, which might lead to lower item difficulty (i.e., result in a higher proportion correct). As part of our efforts to make some items more difficult and more fully span the full range of item difficulties along Dimension RC, we compared the simulated effects of three vs. five response options. In one variant of the five option condition, we held item difficulties the same but reduced the guessing parameter for each five-alternative item to 0.15, to simulate a lower probability of correctly guessing with five options. In the second variant, we posited that more alternatives would increase the difficulty of each item by 0.10 and reduce the guessing to 0.15. For the three-option items, guessing parameters were 0.24, so that reducing the guessing parameter to 0.15 reduced the lower asymptote by 0.09. Five-option items were included in the "ideal" item bank, but not the "real" item bank. In all three levels of this factor, the SEM stopping rule was 0.35 and the upper limit on Phase 2 items was 25.

Dependent Variables

We examined four major dependent variables in Phase 1. The first was the average bias in the estimates of θ along dimension 1:

$$\text{Bias}(\theta) = \frac{1}{N}\sum_{i=1}^{i=N}(\hat{\theta}_{i.RC} - \theta_{i.RC}) \quad (4)$$

which represents the average difference between the estimated location along Dimension RC, $\hat{\theta}_{i.RC}$, and the generated value, $\theta_{i.RC}$, for simulee *i*. We also examined conditional bias (conditional on θ_{RC}). In Equation 4 and subsequent outcome measures, $N = 7{,}500$ for overall statistics and $N = 500$ for outcome measures conditional on θ_{RC}.

The second dependent variable was the root mean square error:

$$\text{RMSE}(\theta) = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}(\hat{\theta}_{i.RC} - \theta_{i.RC})^2} \quad (5)$$

RMSE was also examined conditional on θ_{RC}. Since there is only one RMSE for each level of θ_{RC}, in the ANOVAs described below, the analysis is based on the squared differences $(\hat{\theta}_{i.RC} - \theta_{i.RC})^2$ for each replication within each level of θ_{RC}.

Third, we examined the root mean square SEM Let $\hat{s}(\hat{\theta}_{RC})$ be the estimated standard error for $\hat{\theta}_{RC}$. Our root mean square SEM was defined as:

$$\mathrm{RMS-SEM}(\theta) = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N} \hat{s}^2(\widehat{\theta}_{i.RC})} \quad (6)$$

Finally, we examined the mean number of items administered to the simulees, i.e., average test length.

Results are reported in plots conditional on the RC dimension. Furthermore, a mixed-design ANOVA framework was used to examine differences between the independent variables for the repeated measures. In our design, for Phase 1, the between-subjects variable was true (generated) reading comprehension (RC) ability ($\theta_{i.RC}$), and the within-subjects variable was the manipulated independent variable (e.g., SEM). The dependent variables were bias, root-mean-squared error (RMSE), root mean square of observed SEM (RMS-SEM), and average test length (ATL). Because the sample size can be specified to be arbitrarily large in simulation studies, significance test results are not reported. An unbiased estimate of effect size, omega-squared (ω^2) measures, were computed (see Appendix C for details) and reported. Effect sizes (ω^2) 0.01, 0.06, and 0.14 reflect small, medium, and large effect sizes, respectively (Cohen, 1988, pp. 280–288).

Phase 2

Model

The second response variable for each item was X_{2ij}:

$$\begin{aligned} X_{2ij} &= 1 \text{ if person } i \text{ chose the paraphrase response for item } j \\ &= 0 \text{ if person } i \text{ chose the elaboration response for item } j \\ &= \text{missing if person } i \text{ chose the correct answer} \end{aligned} \quad (7)$$

Note that whether X_{2ij} is defined (i.e., not missing) depends on whether the correct answer has been chosen. To model X_{2ij}, we chose a unidimensional two-parameter logistic (2PL) model. Because X_{2ij} is defined only if $X_{1ij} = 0$, the probability on the left side of this model is conditional on $X_{1ij} = 0$:

$$\pi_{2ij}(X_{2ij} = 1 | X_{1ij} = 0) = \frac{\exp\left[\alpha_{PP.j}(\theta_{i.PP} - \beta_{PP.j})\right]}{1 + \exp\left[\alpha_{PP.j}(\theta_{i.PP} - \beta_{PP.j})\right]} \quad (8)$$

where $\theta_{i.PP}$ is the location of person *i* along Dimension PP, the process propensity dimension, and ($\alpha_{PP.j}$, $\beta_{PP.j}$) is the vector of item parameters (discrimination and difficulty) for Dimension PP and item *j*. In Equation 8, the item index *j* runs over the items taken by student *i* in both Phase 1 and Phase 2. That is, $\theta_{i.pp}$ is estimated from the student's incorrect responses in both phases.

Formulation of the likelihood function requires an assumption of local independence. For this likelihood function, we assume that, for any two items *j* and *j'*, the variables X_{2ij} and $X_{2ij'}$ are independent after conditioning on θ_{PP}, and $X_{1ij} = 0$. This leads to the following likelihood function for the variable 2 response vector of person *i*, $\boldsymbol{x}_{2i} = [X_{2i1}, X_{2i2}, \ldots, X_{2iJ}]$:

$$L_{i.PP} = \prod_{j=1}^{j=J} (\pi_{2ij})^{(1-X_{1ij})X_{2ij}} (1 - \pi_{2ji})^{(1-X_{1ij})(1-X_{2ij})} \quad (9)$$

In Equation 9, the item index *j* runs over all of the items taken in Phase 1 and Phase 2. Equation 9 has the form of the familiar logistic function except that each exponent is a product of $(1 - X_{1ij})$. This likelihood function can be maximized by standard software that can properly handle the missing data of Equation 7. Given that the maximum of the likelihood function is not defined for some response vectors, we again used weighted maximum likelihood estimation.

For the Dimension PP item parameters, we used the real item parameters from the 360 items estimated from the computerized but non-adaptive edition of MOCCA (See Appendix, Table A.1). θ was assumed to take on 15 values from –2.8 to +2.8 in increments of 0.4 along Dimension PP. As with Phase 1, we specified 500 simulees at each of the 15 θ values, resulting in a total of 7,500 simulees. Also as in Phase 1, this uniform distribution of θ_{PP} is not likely to occur in practice, but we chose to simulate θ values in this fashion so we could study how the dependent variables varied conditional on θ.

Independent Variables

Administration: All θ_{RC} vs $\theta_{RC} < 0$. Since the initial conceptualization of MOCCA, we have not reported a process propensity classification for higher achieving students. There were two reasons for this. First, higher ability students have few incorrect answers on which to base the classification. Second, the process propensity classification has been designed for use in individualizing the developmental instruction of poor comprehenders. However, high achieving readers are unlikely to need that type of developmental instruction. But how high should achievement on Dimension RC be before we decide not to proceed to Phase 2? We experimented with no cut-off versus a cut-off at $\theta_{RC} = 0$.

Item Selection Rule: Fisher Information vs. Weighted Fisher Information. Two item selection rule options were investigated. The first option involved choosing the item with the largest Fisher information along Dimension PP conditional on the student's current θ estimate: $I_j(\hat{\theta}_{i.PP})$. For the second option, we evaluated a weighted Fisher information. Because an item yields information about θ_{PP} only if the student incorrectly answered the item, the second option involved weighting the Fisher information along Dimension PP by the probability that the student would incorrectly answer the item:

$$I_j^*(\hat{\theta}_{PP}) = [1 - \pi_{1ij}(X_{1ij} = 1)] \times I_j(\hat{\theta}_{PP}) \quad (10)$$

where $I_j^*(\hat{\theta}_{PP})$ refers to the weighed Fisher information for item *j* at the current estimate of Dimension PP, $\hat{\theta}_{PP}$, for person *i*, $\pi_{1ij}(X_{1ij} = 1)$ refers to the probability that person *i* with Dimension RC estimate $\hat{\theta}_{RC}$ will correctly answer item *j*, and $I_j(\hat{\theta}_{PP})$ is the Fisher information for item *j* along Dimension PP at $\hat{\theta}_{PP}$. This independent variable was a factor with two conditions, Fisher information and weighted Fisher information.

Stopping Rule: Confidence Interval vs. Sequential Probability Ratio vs. Generalized Likelihood Ratio. Three stopping rules were compared: a confidence interval rule, the sequential probability ratio test (SPRT), and the generalized likelihood ratio (GLR) test. As described above, Dimension PP was conceived as a bipolar dimension with elaborating at the negative end and

paraphrasing at the positive end. A person with a paraphrasing process propensity is defined as a person whose probability of choosing the paraphrase response, given that they answer the item incorrectly, is greater than 0.5 for an item of difficulty $\beta_{PP.j} = 0$. This means that the zero point along the dimension divides persons with a paraphrasing process propensity from those with an elaborating process propensity.

For the confidence interval rule, after each item, the algorithm computes the weighted maximum likelihood estimate of the person's Dimension PP location $\hat{\theta}_{PP}$, and the corresponding standard error, $\hat{s}(\hat{\theta}_{PP})$ from the weighted likelihood function (Warm, 1989). From these quantities a 90% confidence interval was computed: CI = $\hat{\theta}_{PP} \pm 1.65\ \hat{s}(\hat{\theta}_{PP})$. If the confidence interval included 0, then the algorithm proceeded to select and administer the next item. If the confidence interval did not include 0, the testing stopped. If the confidence interval was below 0, the person's process propensity was classified as elaborating. If the confidence interval was above 0, the person's process propensity was classified as paraphrasing.

The second stopping rule investigated was the sequential probability ratio test (SPRT; Thompson, Yon, & Berhad (2012); Wang, Chen, & Huebner, 2021). The SPRT begins by establishing an indifference region along Dimension PP about the cut-off separating elaborating from paraphrasing propensities, 0 in our case. Let UB be the upper bound for the indifference region and let LB be the lower bound: LB < 0 < UB. Let $\boldsymbol{X}_i$ be the person's response vector after the j^{th} item is administered. Two likelihoods are computed: the first is the likelihood of $\boldsymbol{X}_i$ at $\hat{\theta}_{PP} =$ UB, and the second is the likelihood at $\hat{\theta}_{PP} =$ LB. Let these two likelihoods be designated as $L(UB|\boldsymbol{X}_i)$ and $L(UB|\boldsymbol{X}_i)$. After the administration of each item, their ratio is computed:

$$LR = \frac{L(\text{UB}|\boldsymbol{X}_i)}{L(\text{LB}|\boldsymbol{X}_i)} \tag{11}$$

Two cut-offs, A and B are then selected such that $0 < A < B$. If LR < A, testing stops, and the person's process propensity is classified as elaborating. If LR > B, then testing stops and the person's process propensity is classified as paraphrasing. If $A <$ LR < B, then testing proceeds to the next item. In our case, we set A = 1/9 and B =9. In the present application, the SPRT will classify a person's process propensity as paraphrasing if the response vector is nine times more likely at the upper bound than at the lower bound. It will classify a person's process propensity as elaborating if the response vector is nine times more likely at the lower bound than at the upper bound.

Our third classification rule, the generalized likelihood ratio test (Wang, Chen & Huebner, 2021) employs three likelihoods, $L(\text{UB}|\boldsymbol{X}_i)$, $L(\text{LB}|\boldsymbol{X}_i)$, and $L(\hat{\theta}_{i.PP}|\boldsymbol{X}_i)$ where $\hat{\theta}_{i.PP}$ is the maximum likelihood estimate of θ_{PP}. Once an item is administered, a new estimate of $\hat{\theta}_{i.PP}$ is obtained. The GLR equals the LR in Equation 11 if LB $< \hat{\theta}_{i.PP} <$ UB. If $\hat{\theta}_{i.PP} \geq$ UB, then the algorithm computes

$$\text{GLR} = \frac{L(\hat{\theta}_{i,PP}|\boldsymbol{X}_i)}{L(\text{LB}|\boldsymbol{X}_i),}. \tag{12}$$

The numerator of the ratio is the θ for which the likelihood is the maximum in the interval $\theta_{i.PP} \geq$ UB, and the denominator is the θ value with the maximum likelihood for values in the range $\theta_{i.PP} \leq$ LB. If $\hat{\theta}_{i.PP} \leq$ LB, the algorithm computes the ratio

$$\text{GLR} = \frac{L(\text{UB}|\boldsymbol{X}_i)}{L(\hat{\theta}_{i.PP}|\boldsymbol{X}_i)}. \quad (13)$$

The quantity in the denominator of the ratio $L(\hat{\theta}_{i.PP}|\boldsymbol{X}_i)$ is the maximum of the likelihood for any value $\theta_{i.PP} \leq$ LB, and the quantity in the numerator is the maximum of the likelihood for any value $\theta_{i.PP} \geq$ UB. Testing will stop, and the person's process propensity will be classified as elaborating, if GLR < A. That is, testing will stop, and the person's process propensity will be classified as elaborating if the maximum of the likelihood in the interval $\hat{\theta}_{i.PP} \leq$ LB is substantially greater than the maximum of the likelihood in the interval $\theta_{i.PP} \geq$ UB. Testing will stop and the person's process propensity will be classified as paraphrasing if GLR > B. That is, the testing will stop, and the person's process propensity will be classified as paraphrasing, if the maximum of the likelihood in the interval $\theta_{i.PP} \geq$ UB is substantially greater than the maximum of the likelihood in the interval $\theta_{I.PP} \leq$ LB. We set A = 1/9 and B = 9, the same values used for the SPRT. In short, the classification stopping rule factor that was studied was a factor with three levels: confidence interval, sequential probability ratio test, and generalized likelihood ratio test.

Stopping Rule: GLR Indifference Regions. Having decided to employ the GLR rule, we briefly studied the effect of varying the "indifference" region, the difference between the upper and lower bounds of the indifference region. We studied the following bounds [–1,00, 1.00], [–0.50, 0.50], and [–0.25, 0.25]. This independent variable was a factor with three levels.

Stopping Rule: Upper Limit on Number of Items. Finally, we experimented with two different stopping rules for the number of items: an upper limit of 25 items for Phase 1 and an upper limit of 15 items for Phase 2 (25+15), versus an upper limit of 25 items for Phase 1 and a total of 40 items for the whole test (25/40). These two stopping rules differ in that, with the first rule, a student can never have more than 15 items in Phase 2, whereas with the second they could have more than 15 items in Phase 2 if they had fewer than 25 items in Phase 1. This independent variable was a factor with two levels here called 25+15 and 25/40.

Dependent Variables

Classification Accuracy. One of the major dependent variables in Phase 2 was classification accuracy. For measuring classification accuracy, simulees with $\theta_{i.PP} > 0$ were classified as having a true paraphrasing propensity, and simulees with $\theta_{i.PP} < 0$ were classified as having a true elaborating propensity. Simulees with $\theta_{i.PP} = 0$ were classified as having neither a true paraphrasing propensity nor a true elaborating propensity; therefore, they were not included in calculations of classification accuracy. We examined the proportion who were correctly classified conditional on true θ_{RC} and θ_{PP}.

Test Length. We were interested in the mean number of items administered in Phase 2.

Other Issues. While we will not present data on these issues below, we also considered two other issues. First we compared two software programs, IRTPRO (Vector Psychometric Group, 2011) and Xcalibre (Guyer & Thompson, 2014) as procedures for fitting the dichotomous data in Phase 1 and Phase 2. Although both programs performed similarly and well for estimation of parameters in Phase 1, Xcalibre yielded more reasonable estimates of parameters in Phase 2 because it permitted us to impose boundaries and moderately tight priors on the estimates. The large amount of missing data generated by the coding of X_{2ij} necessitated the parameter boundaries and the informative priors. Because the coding of X_{2ij} yielded large amounts of missing data, we

also considered a polytomous coding of the process propensity responses in Phase 2 with 3 categories, one for the correct response, one for the paraphrase response, and one for the elaboration response, applying the graded response IRT model. The polytomous coding yielded reasonable results in some respects, but when the polytomous coding was used in Phase 2, the joint distribution of the estimates $(\hat{\theta}_{rc}, \hat{\theta}_{PP})$ did not mirror the shape of the joint distribution for the generating parameters $(\theta_{RC}, \theta_{PP})$ of our simulated data, and therefore we rejected estimation of θ_{PP} based on polytomous coding (See Figure B.1). All results below are based on the dichotomous coding of the Phase 2 variable described above using Xcalibre software.

Results are shown in plots conditional on the RC and PP dimensions. Furthermore, the results were analyzed by mixed-design ANOVA. The between-subjects variables were true (generated) RC θ or true (generated) PP θ, and the within-subjects variable was the manipulated independent variable (e.g., Item selection: Fisher information vs. modified Fisher information). The dependent variables were classification accuracy and Phase 2 average test length. As in the Phase 1 results, effect sizes (ω^2) are reported in favor of significance test results. Because classification accuracy is a binary dependent variable, and the average classification rate in many cells of the design was near 100%, typical ANOVA assumptions were not met. However, there were no suitable alternative models that both converged and allowed effect sizes to be calculated, so ANOVA ω^2s are reported with caution.

RESULTS

Item Banks

For the item parameters, we used the real test item parameters for 360 items developed for the computerized but non-adaptive edition of MOCCA. These parameters are shown in Appendix Table A.1,with overall model fit statistics in Table A.2. For the RC Dimension, the mean and standard deviation of the discrimination parameters were 1.899 and 0.394, respectively, and for the difficulty parameters they were −0.217 and 0.441. Guessing parameters were fixed at 0.24 for all items. For the PP Dimension, the mean and standard deviation of the discrimination parameters were 1.171 and 0.173, while those for the difficulty parameters were −0.351 and 0.558. Figure 2 shows histograms of the discrimination and difficulty parameters for the RC dimension (top) and PP dimension (bottom). The distributions of the difficulty parameters for both dimensions are centered just below zero, while the majority of the difficulty parameters fall within the range of −1 to 1. This narrow distribution of difficulty parameters leads to the peaked bank information functions seen in Figure 3, which also displays the conditional standard error plots for each dimension. The bank is more informative at the peak for the RC dimension than the PP dimension due to the higher discrimination values for the former (shown in Figure 2). However, the lower discrimination values for the PP dimension give its bank information function a broader shape, leading to more information at the extremes when compared to the RC dimension.

Phase 1

Stopping Rule: Estimated Standard Error of Measurement (SEM). Two test termination criteria were evaluated for MOCCA CAT Phase 1: (1) observed SEM, meaning a test will be terminated when a preset SEM (e.g., 0.35) or test length is obtained for a simulee, and (2) maximum test length, whichever comes first. SEM represents the estimated precision of θ_{RC} trait estimates. Lower values represent more restrictive criteria that result in a longer test. The precision and accuracy of Phase 1 θ estimates were compared across two SEM stopping conditions (0.30 and 0.35). Table A.3 shows the results of the ANOVA for SEM, with means for all dependent variables displayed in Figure 4 (means and SDs for Bias, RMSE, RMS-SEM, and test length are in Tables A.10 through A.13, respectively). The ANOVAs resulted in ω^2 values for θ_{RC} that were moderate (0.18 for bias and 0.16 for RMSE) to high (0.47 for RMS-SEM and 0.70 for ATL). All other effects in the ANOVA resulted in $\omega^2 < .001$. As shown in Figure 4, an SEM of 0.30 did not result in markedly more accurate results relative to an SEM of 0.35, despite slightly increasing the average test length conditional on true θ_{RC} levels. Consequently, an SEM of 0.35 was adopted as a Phase 1 termination criterion.

Figure 2. Distributions of α and β Parameters for RC and PP Item Banks

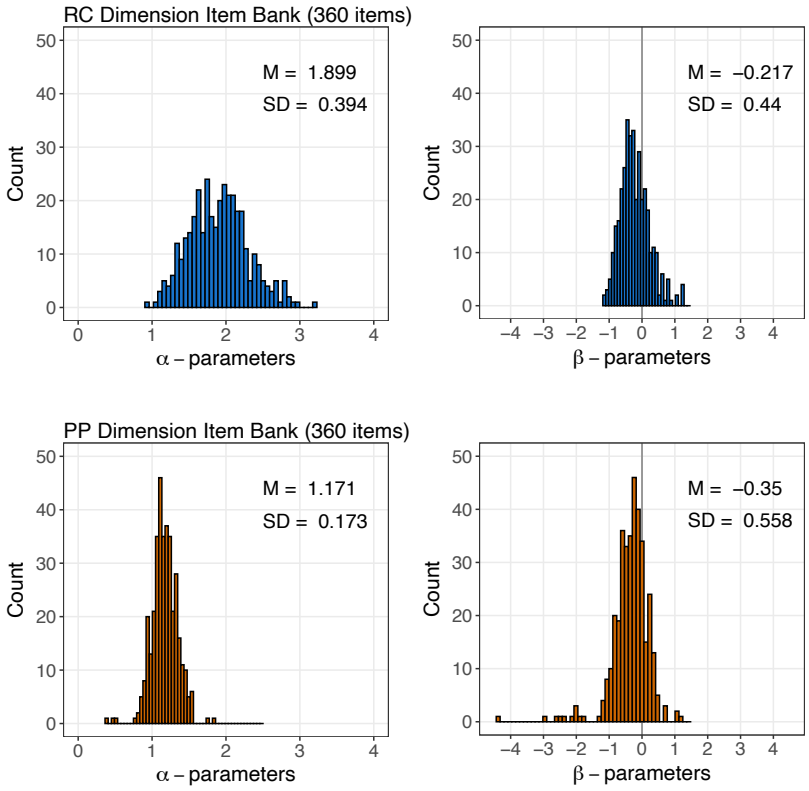


Figure 3. Conditional Information and SEM Functions for RC and PP Item Banks

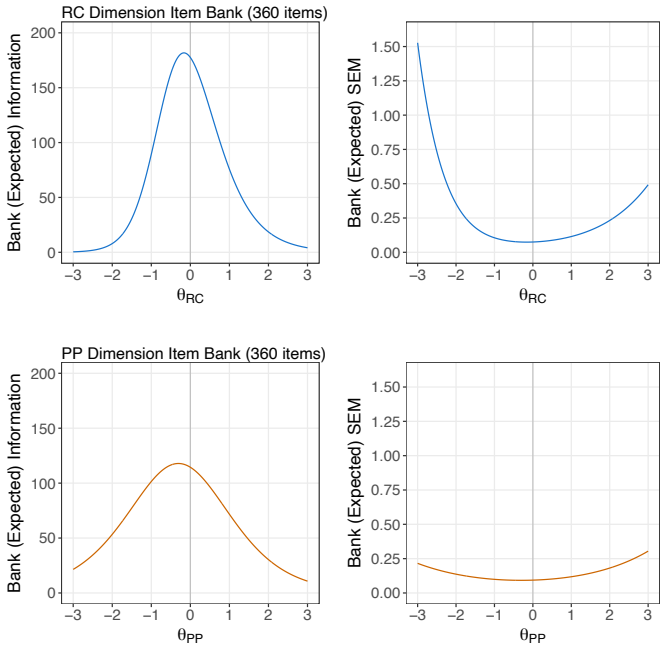
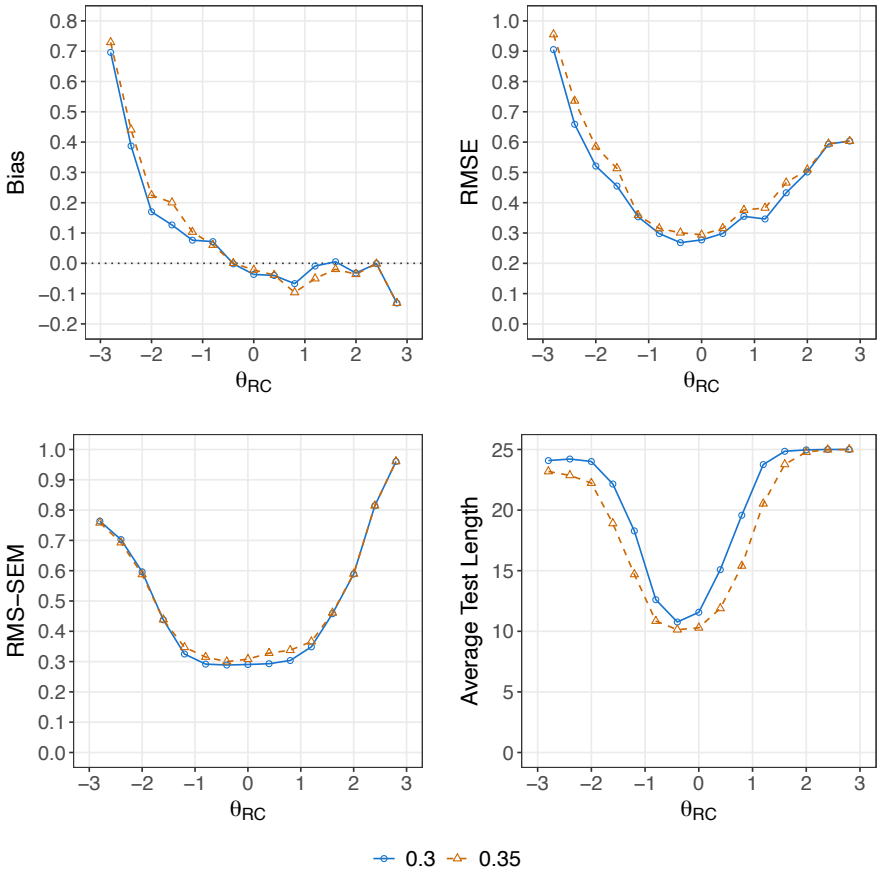


Figure 4. Variation in Dependent Variables as a Function of Two SEM Stopping Criteria (.30 and .35) Conditional on θ_{RC}



Maximum Test Length. Another test termination criterion in MOCCA CAT was the upper bound on the number of items administered (maximum test length, or MTL). The non-adaptive computerized version of MOCCA includes 40 items; therefore, with a goal of decreasing the test length of the CAT, upper bounds of 40, 30, and 25 items were compared. As shown in Table A.4 and Figure 5 (means and SDs are in Tables A.14 – A.17), there was little effect of MTL on bias, RMSE, or RMS-SEM across the three levels of MTL ($\omega^2 < 0.01$ for all dependent variables except average test length, where $\omega^2 = 0.06$). The dependent variables all varied across θ_{RC}, with ω^2 ranging from 0.16 for RMSE to 0.62 for ATL. Figure 5 (bottom panel), shows that an MTL of 25 items produced lower average Phase 1 test length when compared to maximum test lengths of 30 and 40 items for the levels of RC between $\theta = -2.8$ and $\theta = -1.2$, as well as between $\theta = 0.8$ and $\theta = 2.8$, with no effect for θ between these values. Based on these results, the upper bound for the Phase 1 test length was set at 25 items.

Number of Response Options. For 240 additional items in the ideal item bank, we evaluated the effect of three vs. five response options. For items with five response options, the c parameter was reduced to 0.15 from 0.24 (lower guessing effect). In one experimental condition, β parameters remained the same, and in other conditions, β parameters were increased by 0.10 and 0.25.

Figure 5. Variation in Dependent Variables as a Functon of MaximumTest Length (25, 30, and 40 items) Conditional on θ_{RC}

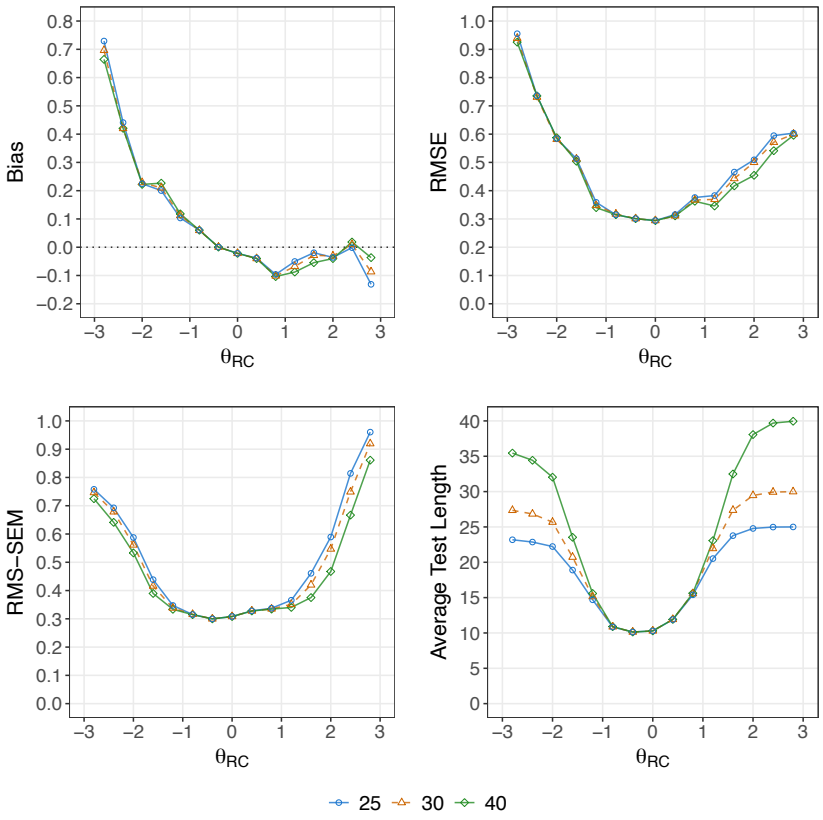


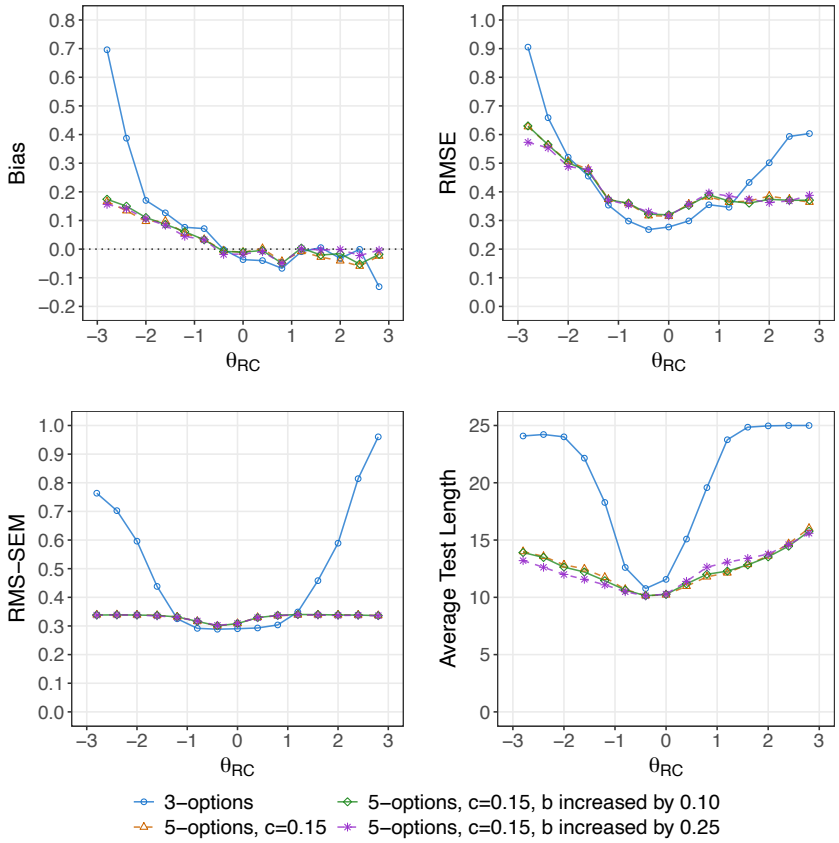
Table A.5 shows that there were only very minor effects for bias and RMSE, with ω^2 ranging from <0.01 to 0.06 for all effects. For RMS-SEM, ω^2 ranged from 0.12 to 0.26, and for ATL it ranged from 0.15 to 0.31, with the largest effect for number of response options. As shown in Figure 6 (and Tables A.18 – A.21) three-alternative items resulted in slightly larger average bias and RMSE, and larger RMS-SEM and ATL. Items with three response options produced a larger Phase 1 average test length when compared to all the other response options between $\theta = -2.8$ and $\theta = -1.2$, as well as between $\theta = 0.4$ and $\theta = 2.8$. All conditions of items with five-alternative response options performed similarly. Based on these results, item writers were requested to write new items with five, rather than three, response alternatives, with one correct answer and two incorrect answers of both types.

Phase 2

Administration: All $\hat{\theta}_{RC}$ vs $\hat{\theta}_{RC} < 0$. Because MOCCA was not intended to classify the process propensity type of high comprehending students, we investigated the effects of a cutoff value for who would be administered Phase 2 of the MOCCA CAT. This cutoff value is based on $\hat{\theta}_{RC}$ at the end of Phase 1. For a cutoff value of 0.0, only simulees with $\hat{\theta}_{RC} < 0$ were administered Phase 2, meaning MOCCA CAT only attempted to classify simulees with below average estimated reading comprehension. Because this cutoff was based on the estimated θ and not the true θ, some

simulees with true $\theta_{RC} > 0$ were still administered Phase 2, and some with $\theta_{RC} < 0$ were not. Table 1 shows the frequency of Phase 2 administration at each true θ_{RC}.

Figure 6. Variation in Dependent Variables as a Function of Number of Response Options (3 Options vs. 5 Options) Conditional on θ_{RC}



We evaluated the classification accuracy and Phase 2 test length of the MOCCA CAT when administering Phase 2 only to examinees with $\hat{\theta}_{RC} < 0$ as compared to administering Phase 2 to all examinees. Classification accuracy was evaluated only with respect to examinees who were administered Phase 2; it is the percentage of attempted classifications that were correct.

Figure 7 shows the results conditional on θ_{RC} (top) and conditional on θ_{PP} (bottom). For the plots of the results conditional on θ_{RC}, a dotted line is used at $\theta_{RC} > 0$ to indicate that these data are from a small sample size due to the $\hat{\theta}_{RC} < 0$ administration rule. For the plots conditional on θ_{PP}, there is no data point for classification accuracy at $\theta_{PP} = 0$, because classification accuracy is not defined when $\theta_{PP} = 0$ — the true classification at $\theta_{PP} = 0$ is neither elaborating nor paraphrasing.

Table 1. Number of Examinees Administered Phase 2 at each True θ_{RC} Under the $\widehat{\theta}_{RC} < 0$ Administration Rule

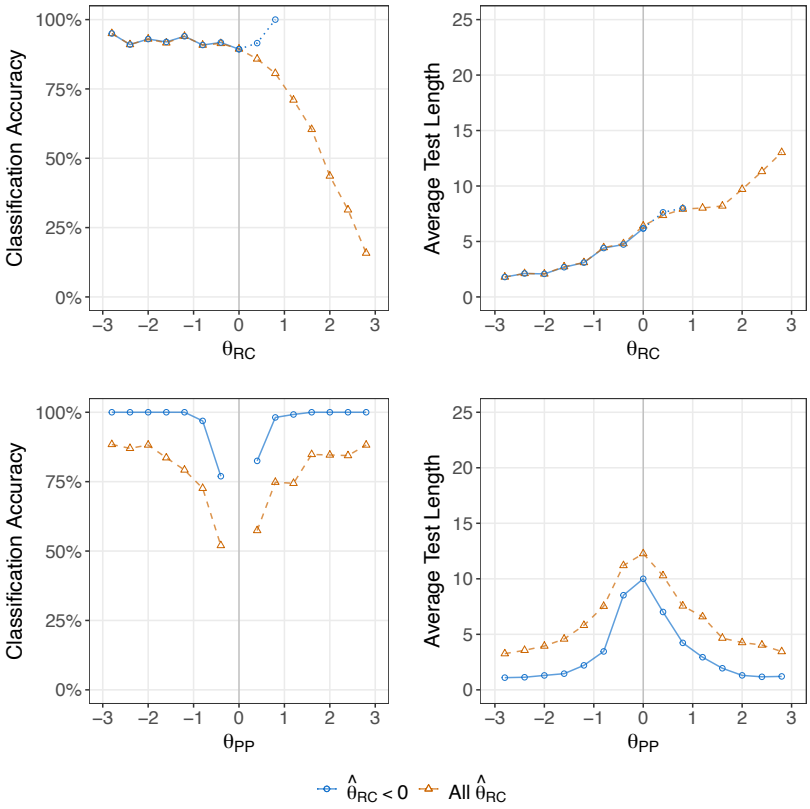
θ_{RC}	*N* Administered Phase 2
-2.8	500
-2.4	499
-2.0	499
-1.6	497
-1.2	499
-0.8	491
-0.4	461
0.0	281
0.4	59
0.8	7
1.2	0
1.6	0
2.0	0
2.4	0
2.8	0

As can be seen in Figure 7, (and Tables A.22 – A.25) the classification accuracy decreased and the test length increased substantially as θ_{RC} increased past zero. By using the $\hat{\theta}_{RC} < 0$ administration rule, we avoid attempting to classify some examinees. Thus, conditional on θ_{PP}, the $\hat{\theta}_{RC} < 0$ administration rule improved the average classification accuracy and test length at all θ_{PP}. This is in part due to the way the data were generated: we generated θ_{RC} and θ_{PP} such that they were uncorrelated, so the impact of the administration rule was evenly spread across θ_{PP}. The result of this rule is that a large portion of simulees were not given a classification at all, but this saves them from taking a longer test when it is both unlikely that they would be classified and unlikely that the classification information would be useful to teachers. Consequently, we adopted the use of the administration rule for the remaining simulations so that only students with $\hat{\theta}_{RC} < 0$ proceed to Phase 2. The same plotting convention of a dotted line at $\theta_{RC} > 0$ is used in the remaining plots of the Phase 2 results.

Item Selection Rule: Fisher Information vs. Weighted Fisher Information. When considering item selection rules, we hypothesized that weighting the Fisher information on θ_{PP} by the probability of an incorrect response (as determined by θ_{RC}) would increase the proportion of incorrect responses during Phase 2, and therefore improve the classification accuracy and test length of the MOCCA CAT. The magnitude of this effect would be dependent upon the underlying θ_{RC} trait value. Simulations based on the full range of θ_{RC} showed this to be the case: both classification accuracy and test length were improved for above average comprehenders.

Table A.6 shows that for classification accuracy conditional on θ_{RC}, there was a negligible main effect, $\omega^2 < 0.01$, and a negligible main effect of item selection rule on classification accuracy, $\omega^2 < 0.01$. Similarly, the interaction effect between θ_{RC} and item selection rule was

Figure 7. Classification Accuracy and Average Test Length as a Function of Including vs. Deleting Cases with $\widehat{\theta}_{RC} < 0$ Conditional on θ_{RC} (top) and θ_{PP} (bottom)

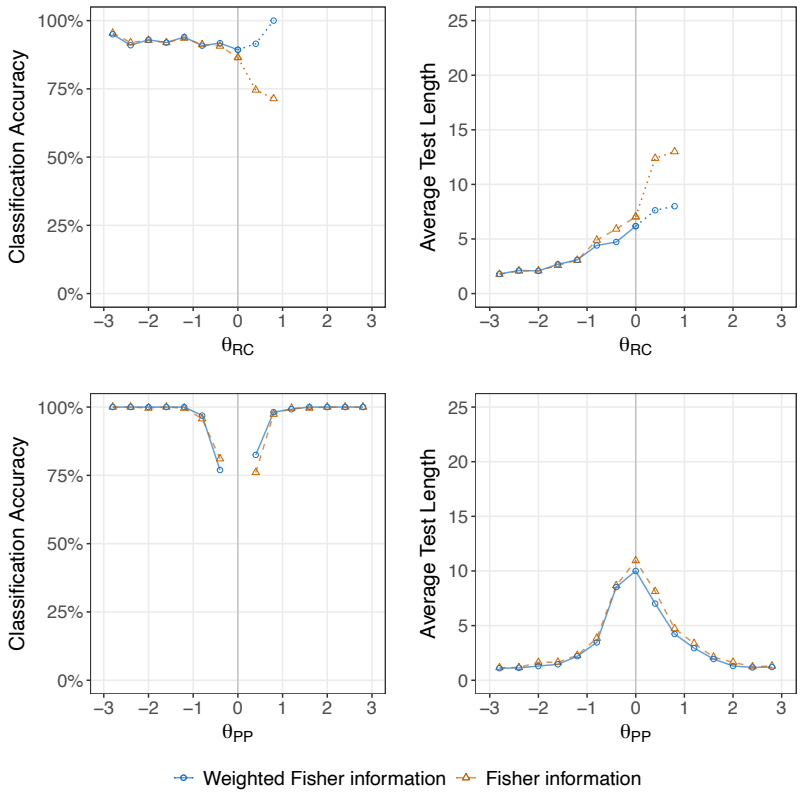


extremely small, $\omega^2 < 0.01$. When classification accuracy was analyzed based on θ_{PP}, there was a main effect of $\omega^2 = 0.44$. There was a negligible main effect of item selection rule on classification accuracy, $\omega^2 < 0.01$. Similarly, the interaction effect between PP and item selection rule was extremely small, $\omega^2 < 0.01$. Average test length conditional on θ_{RC} resulted in a main effect with $\omega^2 = 0.08$, whereas for θ_{PP}, $\omega^2 = 0.23$. All within subjects effects for average test length resulted in very small $\omega^2 < 0.01$.

Figure 8 (see also Tables A.26 – A.29 for means and SDs) shows that after implementing the $\widehat{\theta}_{RC} < 0$ administration rule there was no major effect of weighted Fisher information item selection on the performance of the MOCCA CAT for $\theta_{RC} \leq 0$. The magnitudes of the increase in classification accuracy at $\theta_{RC} > 0$ might not be accurate due to the smaller sample sizes at those points. We nevertheless chose to retain the weighted Fisher information item selection rule due to its theoretical properties and slightly better average test length for $\theta_{RC} = [-0.8, -0.4, 0.0]$.

Stopping Rule: Confidence Interval vs. Sequential Probability Ratio vs. Generalized Likelihood Ratio. MOCCA CAT Phase 2 has two test termination criteria: test length and classification rule, meaning that a test will be terminated when the classification rule is able to classify the examinee's comprehension process as paraphrasing or elaborating, or the test length reaches the maximum allowed, whichever comes first. Examinees who reach the maximum test length are then classified as "inconclusive," meaning the classification rule was unable to classify them as paraphrasing or elaborating because they had an approximately equal mix of paraphrase and elab-

Figure 8. Classification Accuracy and Average Test Length as a Function of Fisher Information vs. Weighted Fisher Information Conditional on θ_{RC} (top) and θ_{PP} (bottom)



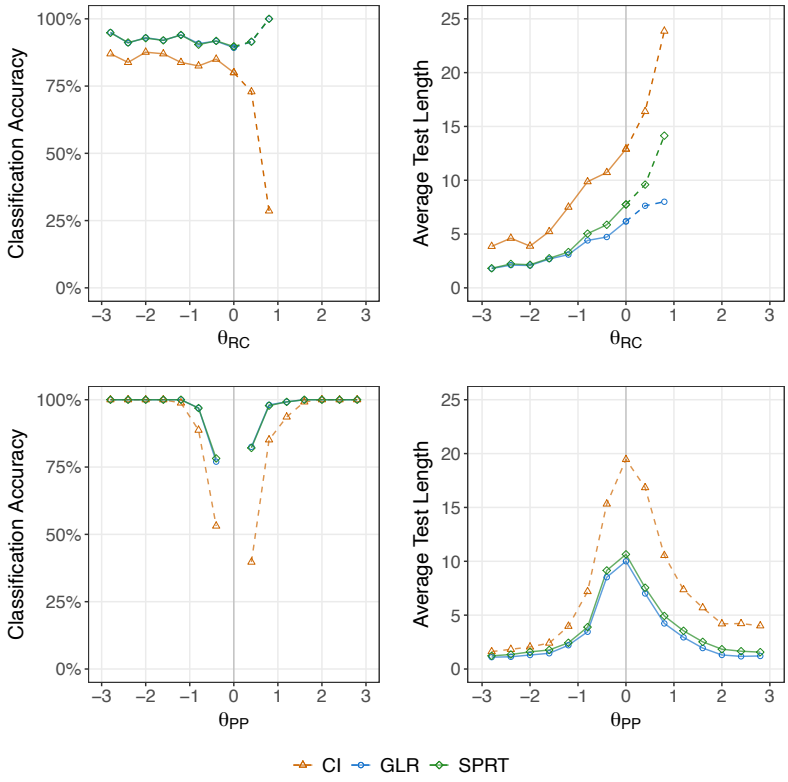
orate responses. Table A.7 shows the results of the ANOVA for the Phase 2 stopping rule based on classification accuracy and test length. For classification accuracy, all ω^2 were 0.01 or less except for θ_{PP}, which had a moderate $\omega^2 = 0.45$. Average test length resulted in $\omega^2 = 0.10$ for θ_{RC} and 0.28 for θ_{PP}. Small effects for stopping rule on test length ($\omega^2 = 0.06$) were obtained for both the RC and PP dimensions and $\omega^2 = 0.03$ for the θ_{PP} × stopping rule interaction.

Figure 9 presents the comparison of the three classification rules (numerical values are in Tables A.30 – A.33. The CI rule had the lowest classification accuracy, while the GLR and SPRT had near identical accuracy. We chose to retain the GLR rule as it was able to achieve the same classification accuracy as SPRT with slightly fewer items administered, and it has a stronger theoretical rationale.

Stopping Rule: GLR Indifference Regions. After deciding to use the GLR classification rule, we then examined the effect of varying the width of the indifference region. The indifference region of the GLR rule affects the strictness of the stopping criteria, with a smaller indifference region representing a more restrictive criterion. Table A.8 shows the results of the ANOVAs for the GLR indifference width variable with dependent variables classification accuracy and ATL.. As the table shows, all effects based on classification accuracy had ω^2 of 0.01 or less, with the exception of θ_{PP}, which resulted in $\omega^2 = 0.44$. A different picture emerged with respect to average test length. There were moderate effects for θ_{RC} ($\omega^2 = 0.05$) as well as for θ_{PP}. There was a large

effect ($\omega^2 = 0.23$) for θ_{PP} and a smaller effect for the interaction of θ_{PP} and indifference region. For the misclassification rate criterion (Table A.8), there was also was a large effect ($\omega^2 = 0.29$) for θ_{PP} and a smaller effect for the interaction of θ_{PP} and indifference region ($\omega^2 = 0.04$). There was a larger effect of the interaction of θ_{PP} and indifference region ($\omega^2 = 0.13$) on the rate of inconclusive classifications, in addition to main effects of θ_{PP} ($\omega^2 = 0.15$) and of indifference region ($\omega^2 = 0.04$).

Figure 9. Classification Accuracy and Average Test Length as a Function of Classification Stopping Rule Conditional on θ_{RC} (top)and θ_{PP} (bottom)



As can be seen in Figure 10 (and Tables A.34 – A.37), the largest effects of the indifference region were near the cut point of 0 on θ_{PP}. We chose to retain the indifference region of [–0.50, 0.50] in order to balance classification accuracy, misclassification rate, and test length. While the indifference region of [–1.00, 1.00] might appear to perform better than the [–0.50, 0.50] region in terms of classification accuracy and test length, it leads to overconfident classifications and an increased misclassification rate for θ_{PP} values near 0 (see Figure 11 and Tables A.38 – A.41).

Figure 30. Classification Accuracy and Average Test Length as a Function of GLR Indifference Region Conditional on θ_{RC} (top) and θ_{PP} (bottom)

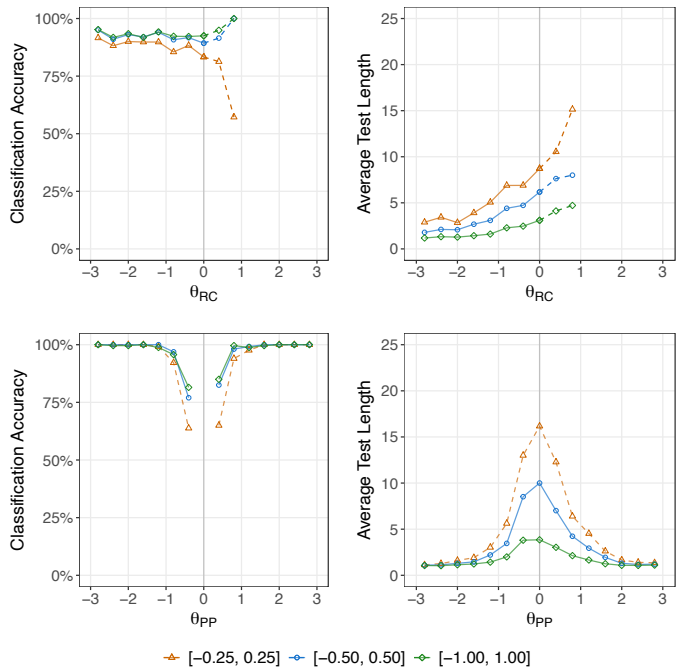
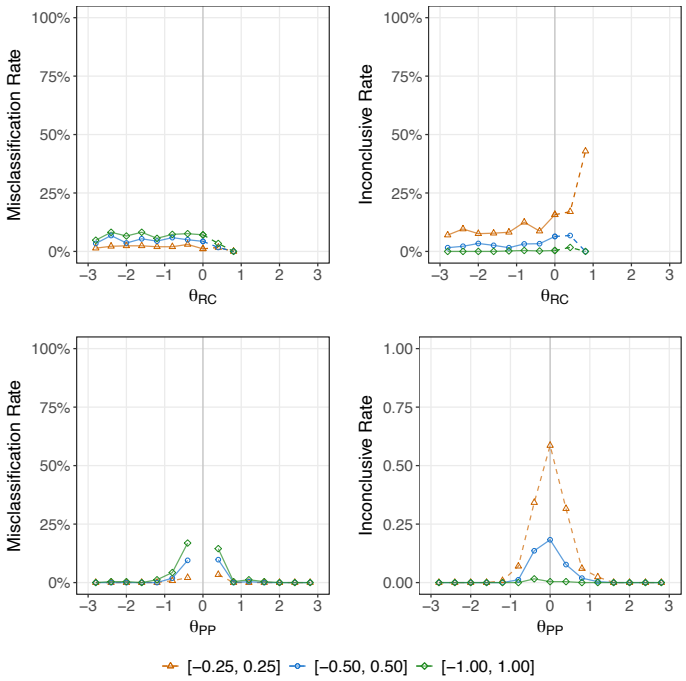


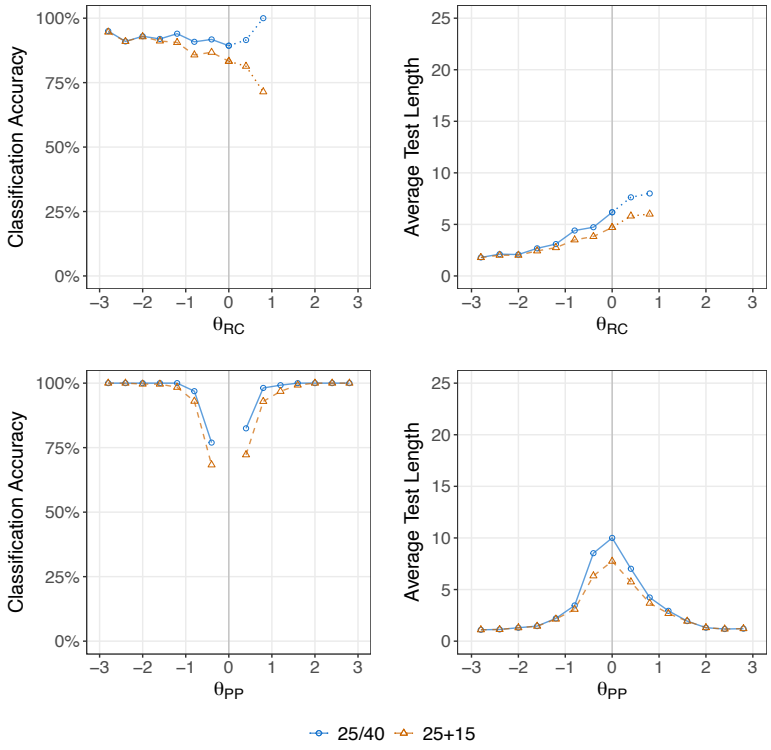
Figure 41. Misclassification Rate and Inconclusive Rate as a Function of GLR Indifference Region Conditional on θ_{RC} (top) and θ_{PP} (bottom)



Stopping Rule: Upper Limit on Number of Items. We considered two options for the upper limit on test length in Phase 2. The first, labeled 25+15, was a hard maximum of 15 items in Phase 2; the second, labeled 25/40, allowed more than 15 items in Phase 2 if fewer than 25 items were administered in Phase 1, as long as the combined number of items administered was 40 or less. Table A.9 shows the ANOVA results for item limit in Phase 2 using classification accuracy and average test length as the dependent variables. All effects resulted in $\omega^2 < 0.01$ except for θ_{PP}, which had $\omega^2 = 0.42$ for classification accuracy, and 0.24 for average test length. θ_{RC} resulted in $\omega^2 = 0.06$

Figure 12 (and Tables A.42 – A.45) shows that because the 25/40 rule allows for the possible administration of more than 15 items in Phase 2, it leads to slightly increased test lengths, especially as θ_{RC} increases and θ_{PP} approaches zero. However, we judged the increase in classification accuracy gained from administering those extra items to be valuable, and so retained the 25/40 rule.

Figure 12. Classification Accuracy and Average Test Length as a Function of Maximum Phase 2 Test Length Conditional on θ_{RC} (top) and θ_{PP} (bottom)



Discussion and Conclusions

This study was conducted to guide decisions about the design of MOCCA CAT. The goal of MOCCA CAT is to develop a test that accurately classifies struggling readers in terms of their propensity to engage in paraphrasing or elaborating processes, if indeed they have such a propensity. We wanted to improve Process Propensity classification without sacrificing measurement accuracy along Dimension 1, the Reading Comprehension dimension, and without increasing test length. Where possible, our goal was to decrease test length. To accomplish these goals, we implemented a sequential, variable-length CAT testing process. In the first phase, students are administered MOCCA items for purpose of precisely locating the student along the Reading Comprehension dimension. Based on the student's Reading Comprehension score, we decide if they might be a reader for whom supplemental instruction could benefit from knowledge of the student's Process Propensity. For good comprehenders, testing ends with Phase 1. Potentially struggling readers proceed to Phase 2 in which they are administered additional items for purposes of determining whether the student has a Process Propensity, and if so, if it is a propensity toward the paraphrasing process or the elaborating process.

Phase 1

In the simulations for the design of Phase 1, we studied three independent variables: stopping rule, upper limit of test length, and number of item options. There were four dependent variables, bias, root mean square error (RMSE), root mean square standard error of measurement (RMS-SEM), and average test length of the variable-length CAT. The first independent variable was a stopping rule with two levels: stop when the student's estimated standard error of measurement falls below 0.30 vs. when the student's estimated standard error falls below 0.35. Within the constraints of our item bank and feasible test lengths, the stricter stopping rule, SEM = 0.30, led to increased test length without much improvement in average bias, RMSE, or RMS-SEM. Once a student reached an estimated SEM $\leq$ 0.35, additional items tended to yield diminishing returns in terms of improved measurement accuracy. Therefore, we have adopted a stopping rule of SEM $\leq$ 0.35.

Our second independent variable was the maximum number of items for Phase 1 with three levels: 25, 30, and 40. Once most students had taken 25 items, additional items yielded diminishing returns in terms of improved bias, RMSE, and RMS-SEM. Therefore we have adopted an upper limit of 25 items in Phase 1.

Our third independent variable was the number of item options: 3 vs. 5. In our simulated data, items with five options and a reduced probability of guessing resulted in better performance of the CAT, particularly with respect to the RMS-SEM and test length. Therefore, as we develop new items, we are including some items with five options.

Phase 2

For Phase 2, we investigated five independent variables: (1) administration of Phase 2 (administration to all students vs administration to below average readers only); (2) two types of item information statistics for selecting items during the CAT (Fisher information vs. weighted Fisher information); (3) three different classification rules [confidence interval rule (CI), sequential probability ratio test (SPRT), and generalized likelihood ratio rule (GLR)]; (4) three widths of

indifference regions for the GLR ([–.25. .25], [–.50, .50], and [–1.00, 1.00]); (5) and two test lengths (25 Phase 1, 15 Phase 2 vs. 25 Phase 1, 40 overall). There were two primary dependent variables: classification accuracy and test length.

Administering Phase 2 only to below average readers led to a substantial reduction in test length for some simulees. It also improved classification accuracy in that there is no attempt to classify readers who are difficult to classify as a having a paraphrasing or elaborating process because they answer few items incorrectly ; students who perform in this manner would be least likely to benefit from Process Propensity information. Therefore, the CAT will identify below average readers as those with estimated Reading Comprehension scores below the mean of our calibration sample, and the CAT will end after Phase I for those whose estimated Reading Comprehension is above the mean.

As compared to Fisher information, weighted Fisher information did lead to better classification accuracy and shorter test lengths, but primarily for simulees with true Reading Comprehension scores above the mean, few of whom will enter Phase 2. Because it did lead to some improvements, however, and is easy to implement, we have decided to use weighted Fisher information in selecting items for the CAT.

In our comparisons of classification rules, the GLR and SPRT rules had better classification rates and shorter test lengths than did the CI rule. There were only minor differences between the SPRT and GLR, but because the GLR displayed slightly better performance and has a stronger theoretical rationale we have decided to implement the GLR in CAT MOCCA.

The width of the indifference region did not yield differences that consistently favored one width over the others across the dependent variables. The indifference region [–.50, .50] seemed to offer a good compromise, and we are adopting it as the indifference region for implementing the GLR for MOCCA CAT.

Finally, for the test length options (25/40 vs. 25+15), the 25/40 resulted in longer test lengths, but higher classification accuracy. In our view the higher accuracy justifies the minor increase in mean length, so we have adopted the 25/40 rule: a limit of 25 items for Phase 1 and a total test limit of 40 items.

Design Summary. Our final CAT design is a variable-length CAT with two phases. In Phase 1, students take MOCCA items to measure their Reading Comprehension. At each step of Phase 1, the item administered is the one with the highest Fisher information for the Reading Comprehension dimension from among the items not yet administered. Testing proceeds until the student's estimated standard error of measurement falls below 0.35 or the number of administered items reaches 25, whichever comes first. If the student's Phase 1 Reading Comprehension score is above the mean for our calibration sample, testing will end with Phase 1. If their Reading Comprehension score is below the mean, the student will proceed to Phase 2. At each step of Phase 2, the student's estimated Process Propensity score will be based on incorrect responses from both Phase 1 and Phase 2. At each step of Phase 2, the administered item is the one with the highest weighted Fisher information for the Process Propensity dimension from among those in the item bank that have not yet been administered.

Testing stops when the student's comprehension process is classified as either Paraphrasing or Elaborating using the generalized likelihood ratio with an indifference region of [–0.5, 0.5] or the student has taken a total of 40 items in Phase 1 and Phase 2 combined. The student's score report will include a Reading Comprehension score and a Process Propensity classification, if the student has been classified in Phase 2, but they will not receive a numeric score for the Process Propensity

dimension. The Process Propensity classification is designed to be a qualitative description useful in the design of future instruction for that student.

References

Biancorosa, G., Kennedy, P. C., Carlson, S. E., Yoon, H.-J., Seipel, B., Liu, B., & Davison, M. L. (2019). Constructing subscores that add validity: A case study identifying students at risk. *Educational and Psychological Measurement, 79(1)*, 65 - 84. DOI 10.1177/0013164418763255

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Lawrence Erlbaum Associates.

Davison, M. L., Biancarosa, G., Seipel, B., Carlson, S. E., Liu, B., & Kennedy, P. (2019). *Administration, interpretation, and technical manual 2019: Multiple-Choice Online Comprehension Assessment* (MOCCA Technical Report MTR-2019-1). Minneapolis, MN: University of Minnesota

De Boeck, P., Chen, H., & Davison, M. L. (2017). Spontaneous and imposed speed of cognitive test responses. *British Journal of Mathematical and Statistical Psychology*, 2017, *70(2)*, 225 – 237

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software, 28*, 2 – 28.

Guyer R. & Thompson, N. A. (2014). User's manual for Xcalibre item response theory calibration software, version 4.2.2 and later. Woodbury, MN; Assessment Systems Corporation.

Holland, P. W. & Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure*. (ETS Program Statistics Report 86-89, RR-86-31). Princeton NJ: ETS. Retrieved from https://doi.org/10.1002/j.2330-8516.1986.tb00186.xdoi.org/10.1002/j.2330-8516.1986.tb00186.x, March 25, 2021.

Kincaid, J. P., Fishburne, L. R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease Formula) for Navy enlisted personnel*. (Research Branch Report No. 8-75). Millington, TN: Naval Air Station Memphis (75). Retrieved from http://library.ucf.edu/Web/purl.asp?dpid=DP0008946

Liu, B., Kennedy, P., Seipel, B., Carlson, S. E., Biancarosa, G., & Davison, M. L. (2019). Can we learn from student mistakes in a reading comprehension assessment? *Journal of Educational Measurement, 56,* 815 – 835. DOI.org/10.1111/jedm.12238

McMaster, K. L., Espin, C. A., & van den Broek, P. (2014). Making connections: Linking cognitive psychology and intervention research to improve comprehension of struggling readers. *Learning Disabilities Research & Practice, 29*(1), 17-24.

McMaster, K. L., van den Broek, P., Espin, C. A., White, M. J., Kendeou, P., Rapp, D. N., Bohn-Gettler, K., Carlson, S. E. (2012). Making the right connections: Differential effects of reading intervention for subgroups of comprehenders. *Learning and Individual Differences, 22*, 100–111.

Partchev, I., & De Boeck. (2012). Can fast and slow intelligence be differentiated. *Intelligence*, *40*, 23 – 32.

Rapp, D. N., van den Broek, P., McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading, 11*, 289–312. doi: 10.1080/10888430701530417.

Su, S. & Davison, M. L. (2019). Improving the predictive validity of reading comprehension using response times of correct responses. *Applied Measurement in Education*, *32*(*2*), 166 – 182. DOI 10.1080/08957347.2019.1577247

Thompson, N., Yon, H., & Berhad, M. (August , 2012). Multiple cutscore classification testing with the generalized likelihood ratio test. Paper presented to the Internnational Association for Computerized Adaptive Testing, Sydney Australia.

Vector Psychometric Group. *IRTPRO guide*, 2011. Retrieved from http://vpgcentral.com, January 15, 2021.

Wang, C., Chen, P., & Huebner, A. (2021). Stopping rules for multi-category computerized classiiation testing. *British Journal of Mathematical and Statistical Psychology*, *74 (2),* 184–202. doi: 10.1111/BMSP.12202

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 54(3), 427–50. http://dx.doi.org/10.1007/bf02294627.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*(4), 473–492.

Appendix A: Supplementary Tables

Table A.1. Item Parameter Estimates for Dimensions RC and PP in the Real Item Bank

		Reading Comprehension			Process Propensity	
Index	Item ID	α	β	c	α	β
1	3	1.737	–0.902	0.24	0.405	–0.92
2	4	1.644	–0.797	0.24	0.500	–0.765
3	5	2.260	–0.369	0.24	0.768	–0.778
4	6	2.090	–0.905	0.24	1.217	–0.134
5	7	1.651	–0.29	0.24	1.094	–0.043
6	8	1.720	–0.838	0.24	1.200	–0.114
7	9	2.045	–0.702	0.24	1.369	–0.416
8	11	2.126	–0.642	0.24	1.181	–0.106
9	12	2.726	–0.787	0.24	1.233	–0.449
10	16	1.966	–0.586	0.24	1.275	0.287
11	17	2.217	–0.987	0.24	1.394	–0.639
12	18	1.182	0.825	0.24	1.213	–0.289
13	20	1.600	–0.475	0.24	1.335	0.048
14	21	1.945	–0.064	0.24	1.071	0.386
15	22	1.895	0.087	0.24	1.223	–0.765
16	23	1.916	–0.42	0.24	1.424	–0.257
17	25	2.212	–0.503	0.24	1.337	–0.537
18	26	2.092	–0.329	0.24	1.114	–0.641
19	27	2.080	–0.647	0.24	1.313	–0.35
20	29	1.778	–0.355	0.24	1.222	0.23
21	30	1.343	0.435	0.24	1.110	–0.224
22	31	2.040	–0.532	0.24	1.187	0.025
23	32	1.408	–0.314	0.24	1.349	1.011
24	33	2.672	–0.614	0.24	1.320	–0.999
25	34	1.887	–0.052	0.24	1.100	0.747
26	35	2.116	–0.479	0.24	1.125	–0.511
27	38	2.165	–0.075	0.24	1.117	–0.119
28	41	2.255	–0.383	0.24	1.176	–0.365
29	43	1.339	–0.001	0.24	1.288	–0.769
30	44	2.446	–0.347	0.24	1.017	0.243
31	45	1.458	0.406	0.24	1.026	–0.302
32	46	2.279	–0.715	0.24	1.312	–0.406
33	48	1.642	–0.553	0.24	1.199	–0.019
34	49	1.517	0.227	0.24	1.063	0.345

- continued on the next page -

Table A.1. Item Parameter Estimates for Dimensions RC and PP in the Real Item Bank

		Reading Comprehension			Process Propensity	
Index	Item ID	α	β	c	α	β
35	50	2.395	–0.399	0.24	1.208	–0.924
36	55	1.459	0.166	0.24	1.397	–1.72
37	56	2.220	–0.423	0.24	1.053	–0.062
38	57	1.944	–0.286	0.24	1.212	–1.245
39	58	2.036	–0.602	0.24	1.216	–0.281
40	59	1.425	–0.644	0.24	1.152	0.304
41	60	1.924	–0.103	0.24	1.140	–1.109
42	62	1.809	–0.059	0.24	0.937	0.099
43	64	2.486	–0.44	0.24	1.242	–0.407
44	68	2.088	–0.429	0.24	1.308	–0.027
45	69	1.735	–0.312	0.24	1.290	–0.566
46	71	1.997	–0.373	0.24	1.041	–0.11
47	72	1.727	0.087	0.24	1.081	–1.026
48	73	1.550	–0.564	0.24	1.257	–0.178
49	76	1.774	–0.124	0.24	1.151	0.727
50	77	2.043	–0.966	0.24	1.471	–0.437
51	80	1.904	–0.763	0.24	1.354	–0.536
52	81	1.804	0.142	0.24	1.172	–0.783
53	82	2.367	–0.809	0.24	1.177	–0.997
54	83	1.487	–0.214	0.24	1.125	–0.265
55	84	2.093	–0.436	0.24	1.071	0.245
56	85	1.496	–0.907	0.24	1.446	–0.5
57	86	1.950	0.072	0.24	1.099	–0.044
58	87	1.967	–0.459	0.24	1.119	–0.256
59	89	1.742	–0.486	0.24	1.142	–0.301
60	92	1.548	0.04	0.24	1.269	–0.246
61	93	2.068	–0.559	0.24	1.113	–0.549
62	94	1.694	–0.328	0.24	1.013	0.06
63	95	2.373	–1.06	0.24	1.227	–0.204
64	96	2.232	–0.315	0.24	1.200	–0.763
65	97	2.110	0.094	0.24	1.006	–0.541
66	98	2.159	–0.733	0.24	1.337	–0.452
67	99	1.468	–0.866	0.24	1.461	–0.271
68	101	1.043	1.199	0.24	1.155	0.042
69	102	1.854	0.141	0.24	0.876	–0.233
70	103	1.531	–0.557	0.24	1.301	–0.09
71	104	2.007	0.341	0.24	1.130	–0.525

- continued on the next page -

Table A.1. Item Parameter Estimates for Dimensions RC and PP in the Real Item Bank

		Reading Comprehension			Process Propensity	
Index	Item ID	α	β	c	α	β
72	106	1.395	0.003	0.24	1.342	–0.017
73	107	1.649	0.289	0.24	0.913	–0.93
74	108	1.704	–0.658	0.24	1.089	–0.152
75	109	2.178	0.106	0.24	0.945	–0.468
76	110	2.454	–0.3	0.24	1.170	–0.594
77	112	2.456	–0.487	0.24	1.158	0.007
78	115	1.689	0.161	0.24	1.091	–0.443
79	116	1.207	0.631	0.24	1.052	–0.13
80	118	1.346	0.798	0.24	1.242	–0.021
81	120	1.568	0.064	0.24	1.154	0.063
82	125	1.579	–0.558	0.24	1.354	0.725
83	126	2.033	–0.062	0.24	1.100	–0.715
84	127	1.836	–0.398	0.24	1.165	0.183
85	128	2.115	–0.287	0.24	1.131	–0.832
86	129	1.126	0.213	0.24	1.180	0.263
87	132	2.342	0.005	0.24	0.961	–0.558
88	133	2.484	–0.285	0.24	0.959	–0.255
89	135	2.415	–0.398	0.24	0.930	–0.185
90	136	2.030	–0.418	0.24	1.149	–0.375
91	137	2.067	–0.452	0.24	1.484	–0.472
92	138	2.330	–0.019	0.24	1.196	–0.798
93	140	2.133	–0.45	0.24	1.353	–0.887
94	141	1.952	–0.607	0.24	1.119	–2.39
95	142	2.490	0.089	0.24	1.113	–0.658
96	143	1.324	–0.333	0.24	0.813	–2.97
97	144	1.908	–0.428	0.24	1.438	0.464
98	145	1.478	–0.582	0.24	1.134	–0.592
99	146	1.355	0.087	0.24	1.325	–0.126
100	147	1.777	–0.251	0.24	1.118	–1.007
101	149	2.138	0.104	0.24	1.162	0.301
102	151	1.876	–0.673	0.24	1.219	0.294
103	153	2.190	–0.705	0.24	1.494	0.132
104	154	2.054	–0.302	0.24	1.308	–0.439
105	155	2.008	0.115	0.24	1.518	–1.002
106	156	1.433	–0.714	0.24	1.333	–0.146
107	157	2.033	0.048	0.24	1.111	–0.163
108	158	1.917	–0.449	0.24	1.212	–0.831

- continued on the next page -

Table A.1. Item Parameter Estimates for Dimensions RC and PP in the Real Item Bank

		Reading Comprehension			Process Propensity	
Index	Item ID	α	β	c	α	β
109	159	1.609	–0.662	0.24	1.330	–0.495
110	160	2.063	0.135	0.24	1.058	–0.024
111	161	1.381	0.361	0.24	0.932	0.207
112	162	1.999	–0.182	0.24	1.007	–0.987
113	165	1.637	–0.124	0.24	0.998	–0.645
114	169	2.189	–0.494	0.24	1.162	–0.331
115	170	2.040	–0.321	0.24	1.180	0.093
116	171	1.851	–0.496	0.24	1.091	0.497
117	174	1.936	–0.353	0.24	1.089	–0.057
118	175	2.293	–0.123	0.24	1.055	–0.072
119	176	1.497	–0.483	0.24	1.237	–0.535
120	177	2.392	–1.015	0.24	1.293	–0.008
121	178	2.132	0.031	0.24	1.003	0.038
122	181	1.965	–0.496	0.24	1.249	–0.605
123	182	2.212	–0.447	0.24	1.277	–0.356
124	183	1.556	–0.341	0.24	1.240	–0.887
125	184	2.137	–0.002	0.24	1.472	–0.702
126	185	1.307	0.486	0.24	0.860	–2.06
127	186	2.107	–0.847	0.24	1.167	0.231
128	188	2.221	–0.215	0.24	1.224	–0.062
129	189	2.472	–0.359	0.24	1.377	–0.404
130	190	1.889	–0.187	0.24	1.074	–0.33
131	192	1.507	–0.012	0.24	1.307	0.265
132	195	1.780	0.358	0.24	0.893	–0.79
133	196	1.378	–0.363	0.24	1.113	0.114
134	198	1.993	–0.978	0.24	1.495	–0.232
135	199	2.174	–0.51	0.24	1.168	–0.622
136	201	1.734	0.426	0.24	0.905	0.09
137	202	2.850	–0.521	0.24	1.439	–0.323
138	203	2.030	–0.488	0.24	1.203	–0.347
139	204	1.524	–0.364	0.24	1.360	–0.265
140	205	1.805	0.068	0.24	1.138	–0.241
141	207	1.362	–0.11	0.24	1.232	–0.326
142	208	1.696	–0.615	0.24	1.175	–1.16
143	211	2.101	–0.681	0.24	1.342	–0.599
144	212	1.436	–0.295	0.24	1.151	–0.32
145	213	1.409	0.667	0.24	1.080	0.378

- continued on the next page -

Table A.1. Item Parameter Estimates for Dimensions RC and PP in the Real Item Bank

		Reading Comprehension			Process Propensity	
Index	Item ID	α	β	c	α	β
146	214	1.282	0.587	0.24	1.165	–1.144
147	215	1.773	0.181	0.24	1.046	–0.333
148	216	2.213	–0.091	0.24	1.122	–0.506
149	217	2.526	–0.48	0.24	1.257	–0.866
150	218	1.951	0.261	0.24	0.899	–0.956
151	219	1.723	–0.204	0.24	1.830	–0.393
152	220	2.225	0.027	0.24	1.254	–0.871
153	221	2.650	–0.771	0.24	1.275	–0.144
154	223	1.856	0.099	0.24	1.192	–0.515
155	225	1.605	0.388	0.24	0.987	–0.176
156	226	1.122	0.663	0.24	1.415	0.352
157	227	1.553	–0.274	0.24	1.001	0.076
158	229	2.581	–0.767	0.24	1.510	–0.18
159	230	1.721	–0.178	0.24	1.054	0.283
160	231	1.554	0.197	0.24	1.103	–0.807
161	232	1.096	–0.261	0.24	1.089	–0.193
162	233	2.182	–0.51	0.24	1.064	–0.353
163	235	1.587	–0.464	0.24	1.146	–0.446
164	236	1.565	0.627	0.24	1.177	–0.652
165	237	1.277	1.279	0.24	0.957	–0.837
166	238	2.243	0.452	0.24	1.000	0.076
167	240	1.561	0.177	0.24	1.060	–0.541
168	241	1.727	–0.013	0.24	1.284	–0.256
169	242	1.739	–0.005	0.24	0.996	–0.075
170	244	2.782	–0.274	0.24	0.948	–0.298
171	245	2.187	–0.5	0.24	1.185	–0.588
172	246	1.608	–0.477	0.24	1.011	0.308
173	247	1.601	–0.321	0.24	0.958	–0.081
174	249	1.725	–0.897	0.24	1.299	–0.448
175	250	1.972	–0.099	0.24	1.074	–0.485
176	251	1.653	–0.083	0.24	1.293	–0.065
177	252	2.664	–0.743	0.24	1.256	–0.029
178	253	2.845	–0.24	0.24	1.328	–1.136
179	254	1.921	0.197	0.24	1.037	–0.387
180	255	1.842	0.355	0.24	1.039	–0.021
181	256	1.503	0.186	0.24	1.413	–0.433
182	258	1.593	–0.309	0.24	1.147	0.399

- continued on the next page -

Table A.1. Item Parameter Estimates for Dimensions RC and PP in the Real Item Bank

		Reading Comprehension			Process Propensity	
Index	Item ID	α	β	c	α	β
183	259	1.722	–0.201	0.24	1.090	–0.044
184	260	2.787	–0.798	0.24	1.099	–0.683
185	262	2.282	0.617	0.24	0.866	–0.527
186	263	1.746	–0.448	0.24	1.172	0.327
187	265	2.056	–0.159	0.24	1.054	–0.529
188	268	1.428	–0.273	0.24	1.231	0.039
189	269	1.323	–0.716	0.24	1.214	–0.451
190	270	1.992	0.424	0.24	0.952	–0.146
191	271	1.624	0.237	0.24	1.101	–0.624
192	272	2.588	–0.748	0.24	1.387	–0.326
193	275	1.718	–0.501	0.24	1.230	–0.587
194	276	1.311	0.166	0.24	1.232	–0.422
195	277	2.058	0.362	0.24	1.209	–0.591
196	278	2.932	–0.649	0.24	1.088	–0.512
197	280	2.583	–0.621	0.24	1.427	–0.185
198	284	1.665	0.759	0.24	1.196	–0.22
199	285	1.815	–0.17	0.24	1.406	0.236
200	286	1.258	–0.01	0.24	1.105	–1.99
201	287	1.500	–0.115	0.24	1.065	1.192
202	288	2.021	–0.618	0.24	1.312	–0.107
203	289	1.780	–0.391	0.24	0.966	–0.17
204	290	2.155	–1.067	0.24	1.319	–0.26
205	291	2.471	–0.551	0.24	1.072	–0.576
206	292	1.607	0.257	0.24	0.894	–0.693
207	293	2.595	0.089	0.24	1.085	0.002
208	294	2.229	–0.001	0.24	0.946	–0.213
209	297	2.394	0.185	0.24	0.950	–0.18
210	299	2.239	–0.064	0.24	0.939	–0.529
211	301	1.927	0.526	0.24	1.038	–0.08
212	302	1.807	–0.039	0.24	1.013	–0.988
213	303	2.332	–0.609	0.24	1.382	–0.154
214	307	1.263	–0.062	0.24	1.098	–1.96
215	308	1.898	0.223	0.24	1.117	0.28
216	309	1.710	–0.795	0.24	1.128	0.052
217	311	1.694	–0.022	0.24	1.084	0.276
218	312	2.174	0.348	0.24	1.110	0.118
219	313	2.051	–0.361	0.24	1.274	–0.56

- continued on the next page -

Table A.1. Item Parameter Estimates for Dimensions RC and PP in the Real Item Bank

		Reading Comprehension			Process Propensity	
Index	Item ID	α	β	c	α	β
220	315	1.946	–0.386	0.24	1.128	–0.133
221	316	1.739	–0.634	0.24	1.304	–0.16
222	318	1.954	–0.103	0.24	0.987	–0.645
223	319	2.358	0.529	0.24	0.929	0.507
224	320	2.152	0.185	0.24	1.136	–1.236
225	321	1.867	–0.291	0.24	1.188	–0.386
226	322	1.511	0.836	0.24	0.970	0.424
227	324	2.237	–0.357	0.24	1.197	–0.675
228	325	2.004	–0.205	0.24	1.118	–0.324
229	326	1.929	–0.265	0.24	1.112	–0.163
230	327	1.567	–0.032	0.24	1.096	–0.204
231	328	2.338	–0.131	0.24	1.027	–0.586
232	329	1.765	0.047	0.24	1.160	–0.208
233	331	1.867	0.129	0.24	1.258	–0.069
234	332	2.001	0.441	0.24	0.826	–0.046
235	333	2.067	–0.555	0.24	1.228	–0.262
236	334	1.834	0.232	0.24	1.314	0.234
237	335	2.003	–0.173	0.24	1.115	0.05
238	336	1.671	–0.113	0.24	1.150	0.133
239	338	1.738	0.204	0.24	1.035	–0.125
240	339	1.572	–0.277	0.24	1.068	–0.115
241	340	1.923	–0.277	0.24	1.320	0.484
242	342	2.029	–0.061	0.24	1.014	–0.655
243	343	2.314	0.226	0.24	1.016	–0.298
244	346	2.142	–0.424	0.24	1.286	–0.641
245	347	1.472	–0.078	0.24	1.113	–0.001
246	349	1.501	–0.115	0.24	1.273	0.207
247	352	1.615	0.028	0.24	1.025	–1.039
248	353	2.051	–0.047	0.24	1.058	–0.122
249	354	1.671	–1.163	0.24	1.540	–0.691
250	355	1.545	–0.472	0.24	1.434	–0.49
251	356	2.297	–0.352	0.24	1.266	–0.506
252	357	2.127	0.426	0.24	1.213	–0.343
253	359	1.582	0.015	0.24	1.220	–0.062
254	361	2.149	–0.082	0.24	1.112	–0.134
255	365	2.292	–0.446	0.24	1.430	–0.272
256	366	2.650	–0.409	0.24	1.111	–0.243

- continued on the next page -

Table A.1. Item Parameter Estimates for Dimensions RC and PP in the Real Item Bank

		Reading Comprehension			Process Propensity	
Index	Item ID	α	β	c	α	β
257	367	2.428	–0.411	0.24	1.222	0.089
258	368	1.810	–0.84	0.24	1.300	–0.055
259	369	2.277	–0.825	0.24	1.313	–0.21
260	370	2.778	–0.222	0.24	1.283	0.359
261	371	1.976	0.094	0.24	0.906	0.218
262	373	2.617	–0.786	0.24	1.559	–0.722
263	374	2.446	–0.173	0.24	1.160	–0.558
264	375	1.734	–0.844	0.24	1.166	–0.695
265	376	1.515	–1.072	0.24	1.552	–0.276
266	378	1.886	1.094	0.24	0.998	–1.094
267	379	1.791	–0.253	0.24	1.069	–0.377
268	382	0.913	–0.221	0.24	1.459	1.017
269	385	2.271	–0.14	0.24	1.102	–0.507
270	386	2.230	–0.357	0.24	1.348	–0.295
271	387	2.359	–0.237	0.24	0.978	–0.798
272	390	2.096	0.064	0.24	1.329	–0.765
273	392	1.650	–0.487	0.24	1.321	–0.603
274	393	1.646	–0.799	0.24	1.251	–0.254
275	394	1.445	–0.115	0.24	1.125	–0.22
276	399	1.164	0.862	0.24	1.210	–1.873
277	400	1.381	0.64	0.24	1.511	0.044
278	401	1.403	–0.268	0.24	1.045	–0.969
279	402	2.013	–0.581	0.24	1.125	0.034
280	403	1.904	–0.606	0.24	1.241	0.022
281	404	1.536	–0.482	0.24	1.244	–0.118
282	405	1.476	–0.047	0.24	1.745	–0.465
283	406	2.195	–0.56	0.24	1.438	–0.806
284	407	1.924	–0.156	0.24	1.368	0.202
285	408	1.624	–0.077	0.24	1.326	–1.012
286	409	2.482	–0.575	0.24	1.190	0.367
287	410	1.482	–0.371	0.24	1.215	0.348
288	412	1.486	–0.209	0.24	1.260	–0.505
289	413	2.084	–0.327	0.24	1.331	–0.81
290	414	2.065	–0.203	0.24	1.094	–0.535
291	416	1.879	–0.747	0.24	1.364	–0.586
292	417	2.657	–0.405	0.24	1.335	–0.64
293	419	1.717	0.176	0.24	0.969	–0.815

- continued on the next page -

Table A.1. Item Parameter Estimates for Dimensions RC and PP in the Real Item Bank

		Reading Comprehension			Process Propensity	
Index	Item ID	α	β	c	α	β
294	420	1.826	–0.401	0.24	1.414	–0.19
295	424	2.083	–0.307	0.24	1.139	–0.778
296	425	1.786	–0.497	0.24	1.152	–0.29
297	426	1.620	0.132	0.24	1.251	0.315
298	427	2.055	–0.418	0.24	1.263	0.191
299	429	1.588	–0.826	0.24	0.884	–2.58
300	430	1.964	–0.733	0.24	1.232	0.057
301	432	2.381	–0.548	0.24	1.170	–0.626
302	434	2.769	–0.508	0.24	1.383	–0.687
303	435	1.633	–0.413	0.24	1.115	0.051
304	436	1.483	–0.281	0.24	1.078	–0.866
305	437	1.690	0.429	0.24	0.922	–1.104
306	438	1.688	–0.134	0.24	1.069	–0.626
307	439	1.176	–0.006	0.24	1.311	–0.277
308	440	1.899	0.058	0.24	1.038	–0.733
309	443	2.094	–0.96	0.24	1.221	–0.235
310	444	1.168	–0.907	0.24	1.161	–0.294
311	445	1.240	1.26	0.24	1.121	–0.012
312	446	2.181	–0.477	0.24	1.066	–0.243
313	447	2.595	–0.304	0.24	1.044	–0.867
314	449	2.369	0.217	0.24	0.955	–0.37
315	450	2.089	–0.156	0.24	1.356	–0.642
316	451	2.780	–0.442	0.24	1.203	0.277
317	452	1.950	–0.482	0.24	1.082	–0.566
318	453	1.878	–0.672	0.24	1.326	–0.164
319	454	1.777	–0.251	0.24	1.189	–0.4
320	456	1.801	0.014	0.24	1.252	–0.211
321	457	1.767	–0.29	0.24	1.220	–0.417
322	458	1.432	1.26	0.24	0.842	–0.298
323	459	2.435	–0.198	0.24	1.139	–0.395
324	460	1.977	1.083	0.24	1.018	–0.064
325	461	2.186	–0.745	0.24	1.276	0.285
326	462	1.982	–0.348	0.24	1.384	–0.386
327	463	1.757	0.348	0.24	0.940	–0.01
328	465	1.342	–0.27	0.24	1.234	0.224
329	466	2.549	–0.62	0.24	1.083	–0.4
330	467	1.980	–0.306	0.24	1.345	–0.162

- continued on the next page -

Table A.1. Item Parameter Estimates for Dimensions RC and PP in the Real Item Bank

		Reading Comprehension			Process Propensity	
Index	Item ID	α	β	c	α	β
331	468	2.967	–0.268	0.24	1.187	–0.376
332	469	1.655	–0.351	0.24	1.207	–0.27
333	470	2.277	–0.637	0.24	1.361	0.036
334	471	1.611	0.752	0.24	1.213	–1.1
335	472	1.398	0.293	0.24	0.971	–0.016
336	473	1.910	–0.23	0.24	1.128	0.21
337	475	1.282	0.457	0.24	1.185	0.241
338	476	2.119	–0.369	0.24	0.454	–4.4
339	477	1.328	–0.055	0.24	1.165	0.129
340	479	1.797	–0.373	0.24	1.114	–0.042
341	480	1.767	0.046	0.24	0.923	–0.203
342	482	1.243	0.281	0.24	1.449	–0.196
343	483	2.107	–0.484	0.24	1.050	–0.684
344	485	2.239	–0.525	0.24	1.410	–0.606
345	486	2.030	–0.664	0.24	1.424	–0.353
346	487	3.183	–0.271	0.24	1.064	–0.08
347	488	1.737	–1.108	0.24	1.192	–0.256
348	489	1.319	–0.71	0.24	1.073	–2.1
349	491	1.616	–0.577	0.24	1.532	0.136
350	493	2.094	–0.802	0.24	1.545	–0.438
351	494	1.948	0.069	0.24	0.935	–0.349
352	495	2.158	–0.107	0.24	0.959	–0.521
353	496	1.197	–0.865	0.24	0.862	–2.52
354	497	1.849	–0.143	0.24	1.246	–0.832
355	498	1.693	0.179	0.24	1.121	–0.562
356	499	1.709	0.369	0.24	0.888	–0.105
357	500	1.632	–0.345	0.24	1.198	–0.119
358	501	1.365	0.341	0.24	1.230	0.136
359	503	1.146	0.376	0.24	1.277	–0.198
360	506	1.832	–0.78	0.24	1.160	–0.481

Note. Item parameters are in the logistic metric (D = 1.0). α = item discrimination, β = item difficulty, and c = lower item asymptote (guessing parameter).

Table A.1. Fit Measures for the 3PL Model of the Reading Comprehension Dimension and the 2PL Moldel of the Bipolar Process Propensity Dimension

Dimension	–2LL	χ^2	df	p
Reading Comprehension	184,958.00	5287.40	4320	<.000
Process Propensity	83743.00	8968.96	4680	<.000

Table A.2. Summary of Mixed Design ANOVA for SEM as a Phase 1 Stopping Rule, and for θ_{RC}

Source	*df*	*MS*	*F*	ω^2
		Bias		
Between subjects				
θ_{RC}	14	48.23	122.1	0.18
Error	7485	0.40		
Within subjects				
SEM	1	0.3739	40.97	<0.01
θ_{RC} × SEM	14	0.2705	29.64	<0.01
Error	7485	0.0091		
		RMSE		
Between subjects				
θ_{RC}	14	43.77	109.1	0.16
Error	7485	0.40		
Within subjects				
SEM	1	3.542	178.98	<0.01
θ_{RC} × SEM	14	0.300	15.15	<0.01
Error	7485	0.020		
		RMS-SEM		
Between subjects				
θ_{RC}	14	36.77	489.6	0.47
Error	7485	0.08		
Within subjects				
SEM	1	0.3689	503.01	<0.01
θ_{RC} × SEM	14	0.0546	74.42	<0.01
Error	7485	0.007		
		Average Test Length		
Between subjects				
θ_{RC}	14	30860	1635	0.70
Error	7485	19		
Within subjects				
SEM	1	11642	3356.6	<0.01
θ_{RC} × SEM	14	485	139.8	<0.01
Error	7485	3		

Table A.3. Summary of Mixed Design ANOVA for Maximum Test Length (MTL) as a Phase 1 Stopping Rule, and for θ_{RC}

Source	*df*	*MS*	*F*	ω^2
		Bias		
Between subjects				
θ_{RC}	14	73.23	122.9	0.18
Error	7485	0.60		
Within subjects				
MTL	2	0.0031	0.247	<0.01
θ_{RC} × MTL	28	0.1604	12.78	<0.01
Error	14970	0.0126		
		RMSE		
Between subjects				
θ_{RC}	14	69.96	105.4	0.16
Error	7485	0.66		
Within subjects				
MTL	2	0.6661	45.33	<0.01
θ_{RC} × MTL	28	0.0709	4.822	<0.01
Error	14970	0.0147		
		RMS-SEM		
Between subjects				
θ_{RC}	14	43.65	478.2	0.44
Error	7485	0.09		
Within subjects				
MTL	2	3.0927	718.13	<0.01
θ_{RC} × MTL	28	0.2384	55.36	<0.01
Error	14970	0.0043		
		Average Test Length		
Between subjects				
θ_{RC}	14	105768	1396	0.62
Error	7485	76		
Within subjects				
MTL	2	75034	16865	0.06
θ_{RC} × MTL	28	4784	1075	0.06
Error	14970	4		

Table A.4. Summary of Mixed Design ANOVA for Number of Response Options for Phase 1, and for θ_{RC}

Source	*df*	*MS*	*F*	ω^2
		Bias		
Between subjects				
θ_{RC}	14	22.605	44.706	0.05
Error	7485	0.506		
Within subjects				
Response options	3	7.552	106.187	0.01
θ_{RC} × Response options	42	3.598	50.592	0.03
Error	22455	0.071		
		RMSE		
Between subjects				
θ_{RC}	14	25.255	40.703	0.06
Error	7485	0.620		
Within subjects				
Response options	3	17.358	378.120	0.01
θ_{RC} × Response options	42	3.027	65.945	0.02
Error	22455	0.046		
		RMS-SEM		
Between subjects				
θ_{RC}	14	5.448	540.702	0.12
Error	7485	0.010		
Within subjects				
Response options	3	41.540	4435.404	0.19
θ_{RC} × Response options	42	4.043	431.701	0.26
Error	22455	0.009		
		Average Test Length		
Between subjects				
θ_{RC}	14	13580.455	1092.505	0.28
Error	7485	12.431		
Within subjects				
Response options	3	71599.566	17788.978	0.31
θ_{RC} × Response options	42	2448.167	608.249	0.15
Error	22455	4.025		

Table A.6. Summary of Mixed Design ANOVA on Classification Accuracy and Average Test Length for Item Selection Rule, and for θ_{RC} and θ_{PP}

Source	*df*	*MS*	*F*	ω^2
	Classification Accuracy			
Between subjects				
θ_{RC}	9	0.39	3.00	<0.01
Error	3783	0.13		
Within subjects				
Item selection rule	1	0.05	3.20	<0.01
θ_{RC} × Item selection rule	9	0.14	9.46	<0.01
Error	3783	0.01		
	Classification Accuracy			
Between subjects				
θ_{PP}	14	17.40	257.21	0.44
Error	3778	0.07		
Within subjects				
Item selection rule	1	0.05	3.17	<0.01
θ_{PP} × Item selection rule	14	0.05	3.62	<0.01
Error	3778	0.02		
	Phase 2 Average Test Length			
Between subjects				
θ_{RC}	9	2589.7	43.08	0.08
Error	3783	60.1		
Within subjects				
Item selection rule	1	202.36	28.74	<0.01
θ_{RC} × Item selection rule	9	115.78	16.44	<0.01
Error	3783	7.04		
	Phase 2 Average Test Length			
Between subjects				
θ_{PP}	14	4724	96.69	0.23
Error	3778	49		
Within subjects				
Item selection rule	1	202.36	27.80	<0.01
θ_{PP} × Item selection rule	14	13.02	1.79	<0.01
Error	3778	7.28		

Table A.7. Summary of Mixed Design ANOVA for Stopping Rule: CI vs. SPRT vs. GLR, for θ_{RC} and θ_{PP}

Source	*df*	*MS*	*F*	ω^2
Classification Accuracy				
Between subjects				
θ_{RC}	9	0.42	1.86	<0.01
Error	3783	0.22		
Within subjects				
Stopping rule	2	7.49	312.82	0.01
θ_{RC} × Stopping rule	18	0.17	7.21	<0.01
Error	7566	0.02		
Classification Accuracy				
Between subjects				
θ_{PP}	14	33.81	336.01	0.45
Error	3778	0.10		
Within subjects				
Stopping rule	2	7.49	381.64	0.01
θ_{PP} × Stopping rule	28	1.29	65.46	0.03
Error	7556	0.02		
Phase 2 Average Test Length				
Between subjects				
θ_{RC}	9	6348	58.91	0.10
Error	3783	108		
Within subjects				
Stopping rule	2	16840	1593.11	0.06
θ_{RC} × Stopping rule	18	414	39.13	0.01
Error	7566	11		
Phase 2 Average Test Length				
Between subjects				
θ_{PP}	14	11624	145.4	0.28
Error	3778	80		
Within subjects				
Stopping rule	2	16840	1906.62	0.06
θ_{PP} × Stopping rule	28	739	83.64	0.03
Error	7556	9		

Table A.8. Summary of Mixed Design ANOVA for Stopping Rule: Width of GLR Indifference Region, for θ_{RC} and θ_{PP}

Source	*df*	*MS*	*F*	ω^2
	Classification Accuracy			
Between subjects				
θ_{RC}	9	0.32	1.63	<0.01
Error	3783	0.20		
Within subjects				
Indifference region	2	2.28	116.40	<0.01
θ_{RC} × Indifference region	18	0.09	4.68	<0.01
Error	7566	0.02		
	Classification Accuracy			
Between subjects				
θ_{PP}	14	28.42	300.02	0.44
Error	3778	0.09		
Within subjects				
Indifference region	2	2.28	125.88	<0.01
θ_{PP} × Indifference region	28	0.46	25.63	0.01
Error	7556	0.02		
	Phase 2 Average Test Length			
Between subjects				
θ_{RC}	9	2429.7	33.89	0.05
Error	3783	71.7		
Within subjects				
Indifference region	2	9458	760.26	0.05
θ_{RC} × Indifference region	18	215	17.31	<0.01
Error	7566	12		
	Phase 2 Average Test Length			
Between subjects				
θ_{PP}	14	6649	125.6	0.23
Error	3778	53		
Within subjects				
Indifference region	2	9458	1021.8	0.05
θ_{PP} × Indifference region	28	1002	108.3	0.07
Error	7556	9		

Table A.8 (continued). Summary of Mixed Design ANOVA for Stopping Rule: Width of GLR Indifference Region, for θ_{RC} and θ_{PP}

Source	*df*	*MS*	*F*	ω^2
		Misclassification Rate		
Between subjects				
θ_{RC}	9	0.10	1.01	<0.01
Error	3783	0.10		
Within subjects				
Indifference region	2	2.15	134.84	<0.01
θ_{RC} × Indifference region	18	0.02	1.31	<0.01
Error	7566	0.02		
		Misclassification Rate		
Between subjects				
θ_{PP}	14	10.52	175.46	0.29
Error	3778	0.06		
Within subjects				
Indifference region	2	2.15	158.71	<0.01
θ_{PP} × Indifference region	28	0.67	49.17	0.04
Error	7556	0.01		
		Inconclusive Classification Rate		
Between subjects				
θ_{RC}	9	0.24	4.14	<0.01
Error	3783	0.06		
Within subjects				
Indifference region	2	8.68	303.71	0.04
θ_{RC} × Indifference region	18	0.10	3.64	<0.01
Error	7566	0.03		
		Inconclusive Classification Rate		
Between subjects				
θ_{PP}	14	4.84	120.33	0.15
Error	3778	0.04		
Within subjects				
Indifference region	2	8.68	410.12	0.04
θ_{PP} × Indifference region	28	2.08	98.19	0.13
Error	7556	0.02		

Table A.9. Summary of Mixed Design ANOVA for Stopping Rule: Upper Limit on Number of Items, for θ_{RC} and θ_{PP}

Source	*df*	*MS*	*F*	ω^2
		Classification Accuracy		
Between subjects				
θ_{RC}	9	0.50	3.34	<0.01
Error	3783	0.15		
Within subjects				
Item limit	1	1.24	102.28	<0.01
θ_{RC} × Item limit	9	0.15	12.70	<0.01
Error	3783	0.01		
		Classification Accuracy		
Between subjects				
θ_{PP}	14	18.61	224.91	0.42
Error	3778	0.08		
Within subjects				
Item limit	1	1.24	103.38	<0.01
θ_{PP} × Item limit	14	0.14	11.50	<0.01
Error	3778	0.01		
		Phase 2 Average Test Length		
Between subjects				
θ_{RC}	9	1226.7	27.95	0.06
Error	3783	43.9		
Within subjects				
Item limit	1	419.5	160.8	<0.01
θ_{RC} × Item limit	9	48.3	18.5	<0.01
Error	3783	2.6		
		Phase 2 Average Test Length		
Between subjects				
θ_{PP}	14	3290	94.88	0.24
Error	3778	35		
Within subjects				
Item limit	1	419.5	172.6	<0.01
θ_{PP} × Item limit	14	80.2	33.0	<0.01
Error	3778	2.4		

Table A.10. Mean and SD for Bias Conditional on θ_{RC} by SEM as a Phase 1 Stopping Rule

	SEM = 0.3		SEM = 0.35	
θ_{RC}	Mean	SD	Mean	SD
–2.8	0.7	0.58	0.73	0.62
–2.4	0.39	0.53	0.44	0.59
–2	0.17	0.49	0.22	0.54
–1.6	0.13	0.44	0.2	0.47
–1.2	0.08	0.35	0.1	0.34
–0.8	0.07	0.29	0.06	0.31
–0.4	0	0.27	0	0.3
0	–0.04	0.28	–0.02	0.29
0.4	–0.04	0.3	–0.04	0.31
0.8	–0.07	0.35	–0.1	0.36
1.2	–0.01	0.35	–0.05	0.38
1.6	0.01	0.43	–0.02	0.47
2	–0.03	0.5	–0.04	0.51
2.4	0	0.59	0	0.6
2.8	–0.13	0.59	–0.13	0.59
Total	0.08	0.48	0.09	0.51

Table A.11. Mean and SD for RMSE Conditional on θ_{RC} by SEM as a Phase 1 Stopping Rule

	SEM = 0.3		SEM = 0.35	
θ_{RC}	Mean	SD	Mean	SD
–2.8	0.82	1.1	0.91	1.22
–2.4	0.43	0.68	0.54	0.83
–2	0.27	0.44	0.34	0.5
–1.6	0.21	0.33	0.26	0.39
–1.2	0.13	0.19	0.13	0.22
–0.8	0.09	0.17	0.1	0.19
–0.4	0.07	0.1	0.09	0.14
0	0.08	0.1	0.09	0.12
0.4	0.09	0.16	0.1	0.17
0.8	0.13	0.2	0.14	0.21
1.2	0.12	0.17	0.15	0.2
1.6	0.19	0.44	0.22	0.46
2	0.25	0.45	0.26	0.46
2.4	0.35	0.41	0.35	0.41
2.8	0.36	0.32	0.36	0.32
Total	0.24	0.47	0.27	0.53

Table A.12. Mean and SD for RMS-SEM Conditional on θ_{RC} by SEM as a Phase 1 Stopping Rule

	SEM = 0.3		SEM = 0.35	
θ_{RC}	Mean	SD	Mean	SD
–2.8	0.71	0.29	0.7	0.29
–2.4	0.65	0.27	0.64	0.27
–2	0.55	0.23	0.54	0.23
–1.6	0.41	0.16	0.41	0.15
–1.2	0.32	0.07	0.34	0.06
–0.8	0.29	0.02	0.31	0.03
–0.4	0.29	0.01	0.3	0.02
0	0.29	0.01	0.31	0.02
0.4	0.29	0.01	0.33	0.02
0.8	0.3	0.03	0.34	0.02
1.2	0.34	0.06	0.36	0.05
1.6	0.43	0.15	0.44	0.15
2	0.55	0.22	0.55	0.22
2.4	0.74	0.34	0.74	0.34
2.8	0.88	0.37	0.88	0.37
Total	0.47	0.27	0.48	0.26

Table A.13. Mean and SD for Test Length Conditional on θ_{RC} by SEM as a Phase 1 Stopping Rule

	SEM = 0.3		SEM = 0.35	
θ_{RC}	Mean	SD	Mean	SD
–2.8	24.1	3.17	23.2	4.56
–2.4	24.2	2.81	22.9	4.76
–2	24	3.13	22.2	5.1
–1.6	22.1	4.71	18.9	6.15
–1.2	18.3	5.64	14.7	5.31
–0.8	12.6	3.18	10.8	1.92
–0.4	10.8	1.41	10.1	0.66
0	11.6	2.12	10.3	0.89
0.4	15.1	3.52	11.9	2.66
0.8	19.6	4.33	15.4	4.49
1.2	23.8	2.46	20.5	4.58
1.6	24.9	0.88	23.8	2.89
2	25	0.45	24.8	1.25
2.4	25	0	25	0.3
2.8	25	0	25	0
Total	20.39	6	18.63	6.7

Table A.14. Mean and SD for Bias Conditional on θ_{RC} by Maximum Test Length as a Phase 1 Stopping Rule

	25 items		30 items		40 items	
θ_{RC}	Mean	SD	Mean	SD	Mean	SD
–2.8	0.73	0.62	0.7	0.63	0.66	0.64
–2.4	0.44	0.59	0.42	0.6	0.42	0.6
–2	0.22	0.54	0.23	0.54	0.22	0.54
–1.6	0.2	0.47	0.21	0.47	0.23	0.45
–1.2	0.1	0.34	0.11	0.33	0.12	0.32
–0.8	0.06	0.31	0.06	0.31	0.06	0.31
–0.4	0	0.3	0	0.3	0	0.3
0	–0.02	0.29	–0.02	0.29	–0.02	0.29
0.4	–0.04	0.31	–0.04	0.31	–0.04	0.31
0.8	–0.1	0.36	–0.1	0.35	–0.1	0.35
1.2	–0.05	0.38	–0.07	0.36	–0.09	0.33
1.6	–0.02	0.47	–0.03	0.44	–0.05	0.41
2	–0.04	0.51	–0.03	0.5	–0.04	0.45
2.4	0	0.6	0.01	0.57	0.02	0.54
2.8	–0.13	0.59	–0.09	0.59	–0.04	0.6
Total	0.09	0.51	0.09	0.5	0.09	0.49

Table A.15. Mean and SD for RMSE Conditional on θ_{RC} by Maximum Test Length as a Phase1 Stopping Rule

	25 items		30 items		40 items	
θ_{RC}	Mean	SD	Mean	SD	Mean	SD
–2.8	0.91	1.22	0.88	1.23	0.86	1.24
–2.4	0.54	0.83	0.53	0.84	0.54	0.84
–2	0.34	0.5	0.34	0.49	0.35	0.49
–1.6	0.26	0.39	0.26	0.39	0.25	0.38
–1.2	0.13	0.22	0.12	0.21	0.12	0.21
–0.8	0.1	0.19	0.1	0.2	0.1	0.19
–0.4	0.09	0.14	0.09	0.14	0.09	0.14
0	0.09	0.12	0.09	0.12	0.09	0.12
0.4	0.1	0.17	0.1	0.15	0.1	0.15
0.8	0.14	0.21	0.13	0.19	0.13	0.18
1.2	0.15	0.2	0.14	0.2	0.12	0.18
1.6	0.22	0.46	0.2	0.38	0.17	0.28
2	0.26	0.46	0.25	0.49	0.21	0.39
2.4	0.35	0.41	0.33	0.44	0.29	0.48
2.8	0.36	0.32	0.36	0.31	0.35	0.35
Total	0.27	0.53	0.26	0.53	0.25	0.52

Table A.16. Mean and SD for RMS-SEM Conditional on θ_{RC} by Maximum Test Length as a Phase 1 Stopping Rule

	25 items		30 items		40 items	
θ_{RC}	Mean	SD	Mean	SD	Mean	SD
–2.8	0.7	0.29	0.69	0.28	0.67	0.28
–2.4	0.64	0.27	0.62	0.27	0.59	0.25
–2	0.54	0.23	0.52	0.21	0.49	0.2
–1.6	0.41	0.15	0.4	0.13	0.38	0.11
–1.2	0.34	0.06	0.34	0.04	0.33	0.03
–0.8	0.31	0.03	0.31	0.03	0.31	0.02
–0.4	0.3	0.02	0.3	0.02	0.3	0.02
0	0.31	0.02	0.31	0.02	0.31	0.02
0.4	0.33	0.02	0.33	0.02	0.33	0.02
0.8	0.34	0.02	0.33	0.01	0.33	0.01
1.2	0.36	0.05	0.35	0.03	0.34	0.02
1.6	0.44	0.15	0.41	0.11	0.37	0.06
2	0.55	0.22	0.51	0.21	0.44	0.15
2.4	0.74	0.34	0.68	0.31	0.61	0.28
2.8	0.88	0.37	0.84	0.37	0.78	0.36
Total	0.48	0.26	0.46	0.25	0.44	0.22

Table A.17. Mean and SD for Test Length Conditional on θ_{RC} by Maximum Test Length as a Phase 1 Stopping Rule

	25 items		30 items		40 items	
θ_{RC}	Mean	SD	Mean	SD	Mean	SD
–2.8	23.2	4.56	27.4	6.27	35.5	9.87
–2.4	22.9	4.76	26.8	6.61	34.4	10.5
–2	22.2	5.1	25.7	7.09	32	11.3
–1.6	18.9	6.15	20.7	8.11	23.5	11.7
–1.2	14.7	5.31	15.1	6.26	15.6	7.5
–0.8	10.8	1.92	10.9	2.09	10.9	2.44
–0.4	10.1	0.66	10.1	0.66	10.1	0.66
0	10.3	0.89	10.3	0.89	10.3	0.89
0.4	11.9	2.66	11.9	2.8	11.9	2.8
0.8	15.4	4.49	15.6	4.91	15.6	5.1
1.2	20.5	4.58	21.9	6.18	23.1	7.98
1.6	23.8	2.89	27.4	4.66	32.5	8.35
2	24.8	1.25	29.5	2.23	38.1	4.95
2.4	25	0.3	29.9	0.69	39.7	1.95
2.8	25	0	30	0.18	40	0.7
Total	18.63	6.7	20.88	8.92	24.88	13.25

Table A.18. Mean and SD for Bias Conditional on θ_{RC} by Number of Response Options

	3 options		5 options		5 options, b + 0.1		5 options, b + 0.25	
θ_{RC}	Mean	SD	Mean	SD	Mean	SD	Mean	SD
–2.8	0.73	0.62	0.17	0.61	0.17	0.61	0.16	0.55
–2.4	0.44	0.59	0.13	0.55	0.15	0.54	0.14	0.54
–2	0.22	0.54	0.1	0.5	0.11	0.49	0.11	0.48
–1.6	0.2	0.47	0.1	0.47	0.09	0.46	0.08	0.47
–1.2	0.1	0.34	0.06	0.37	0.06	0.37	0.04	0.37
–0.8	0.06	0.31	0.03	0.36	0.03	0.36	0.03	0.35
–0.4	0	0.3	–0.01	0.32	–0.01	0.32	–0.02	0.33
0	–0.02	0.29	–0.01	0.32	–0.01	0.32	–0.02	0.32
0.4	–0.04	0.31	0	0.36	0	0.35	–0.01	0.36
0.8	–0.1	0.36	–0.04	0.38	–0.05	0.39	–0.05	0.39
1.2	–0.05	0.38	–0.01	0.36	0	0.37	0	0.39
1.6	–0.02	0.47	–0.03	0.37	–0.02	0.36	0	0.37
2	–0.04	0.51	–0.04	0.38	–0.02	0.37	0	0.36
2.4	0	0.6	–0.06	0.37	–0.05	0.37	–0.02	0.37
2.8	–0.13	0.59	–0.02	0.36	–0.02	0.37	0	0.39
Total	0.09	0.51	0.02	0.42	0.03	0.42	0.03	0.41

Table A.19. Mean and SD for RMSE Conditional on θ_{RC} by Number of Response Options

	3 options		5 options		5 options, b + 0.1		5 options, b + 0.25	
θ_{RC}	Mean	SD	Mean	SD	Mean	SD	Mean	SD
–2.8	0.91	1.22	0.39	1.1	0.4	1.09	0.33	1.01
–2.4	0.54	0.83	0.32	0.77	0.32	0.76	0.31	0.77
–2	0.34	0.5	0.26	0.52	0.25	0.49	0.24	0.49
–1.6	0.26	0.39	0.23	0.4	0.22	0.38	0.23	0.39
–1.2	0.13	0.22	0.14	0.22	0.14	0.21	0.14	0.21
–0.8	0.1	0.19	0.13	0.27	0.13	0.27	0.13	0.27
–0.4	0.09	0.14	0.1	0.18	0.1	0.19	0.11	0.19
0	0.09	0.12	0.1	0.16	0.1	0.18	0.1	0.17
0.4	0.1	0.17	0.13	0.2	0.12	0.2	0.13	0.22
0.8	0.14	0.21	0.15	0.19	0.15	0.2	0.16	0.22
1.2	0.15	0.2	0.13	0.2	0.14	0.2	0.15	0.22
1.6	0.22	0.46	0.14	0.25	0.13	0.22	0.14	0.24
2	0.26	0.46	0.15	0.23	0.14	0.24	0.13	0.22
2.4	0.35	0.41	0.14	0.19	0.14	0.19	0.14	0.18
2.8	0.36	0.32	0.13	0.2	0.14	0.21	0.15	0.22
Total	0.27	0.53	0.18	0.43	0.17	0.43	0.17	0.41

Table A.20. Mean and SD for RMS-SEM Conditional on θ_{RC} by Number of Response Options

	3 options		5 options		5 options, b + 0.1		5 options, b + 0.25	
θ_{RC}	Mean	SD	Mean	SD	Mean	SD	Mean	SD
–2.8	0.7	0.29	0.34	0.01	0.34	0.01	0.34	0.01
–2.4	0.64	0.27	0.34	0.01	0.34	0.01	0.34	0.01
–2	0.54	0.23	0.34	0.01	0.34	0.01	0.34	0.01
–1.6	0.41	0.15	0.34	0.01	0.34	0.01	0.33	0.01
–1.2	0.34	0.06	0.33	0.02	0.33	0.02	0.33	0.02
–0.8	0.31	0.03	0.32	0.02	0.32	0.02	0.32	0.02
–0.4	0.3	0.02	0.3	0.02	0.3	0.02	0.3	0.02
0	0.31	0.02	0.31	0.02	0.31	0.02	0.31	0.02
0.4	0.33	0.02	0.33	0.02	0.33	0.02	0.33	0.02
0.8	0.34	0.02	0.34	0.01	0.34	0.01	0.34	0.01
1.2	0.36	0.05	0.34	0.01	0.34	0.01	0.34	0.01
1.6	0.44	0.15	0.34	0.01	0.34	0.01	0.34	0.01
2	0.55	0.22	0.34	0.01	0.34	0.01	0.34	0.01
2.4	0.74	0.34	0.34	0.01	0.34	0.01	0.34	0.01
2.8	0.88	0.37	0.34	0.01	0.34	0.01	0.34	0.01
Total	0.48	0.26	0.33	0.02	0.33	0.02	0.33	0.02

Table A.21. Mean and SD for Test Length Conditional on θ_{RC} by Number of Response Options

	3 options		5 options		5 options, b + 0.1		5 options, b + 0.25	
θ_{RC}	Mean	SD	Mean	SD	Mean	SD	Mean	SD
–2.8	23.2	4.56	14	3.15	13.9	3.2	13.2	3.26
–2.4	22.9	4.76	13.6	3.06	13.5	2.98	12.6	2.97
–2	22.2	5.1	12.8	2.78	12.6	2.74	12	2.61
–1.6	18.9	6.15	12.5	2.71	12.2	2.6	11.6	2.29
–1.2	14.7	5.31	11.8	2.19	11.4	2.07	11.1	1.84
–0.8	10.8	1.92	10.7	1.34	10.6	1.3	10.5	1.14
–0.4	10.1	0.66	10.1	0.52	10.1	0.55	10.1	0.55
0	10.3	0.89	10.2	0.66	10.2	0.68	10.3	0.82
0.4	11.9	2.66	11	1.31	11.2	1.38	11.4	1.64
0.8	15.4	4.49	11.8	1.57	12	1.58	12.6	1.85
1.2	20.5	4.58	12.1	1.39	12.3	1.43	13.1	1.44
1.6	23.8	2.89	12.9	1.51	12.8	1.63	13.4	1.67
2	24.8	1.25	13.6	1.51	13.5	1.49	13.8	1.45
2.4	25	0.3	14.7	1.61	14.5	1.63	14.6	1.62
2.8	25	0	16	1.56	15.8	1.57	15.6	1.55
Total	18.63	6.7	12.52	2.54	12.45	2.48	12.39	2.46

Table A.22. Mean and SD for Classification Accuracy Conditional on θ_{RC} by Administration Method: All $\widehat{\boldsymbol{\theta}}_{\boldsymbol{RC}}$ vs $\widehat{\boldsymbol{\theta}}_{\boldsymbol{RC}} < \mathbf{0}$

	All $\hat{\theta}_{RC}$		$\hat{\theta}_{RC} < 0$	
θ_{RC}	Mean	SD	Mean	SD
–2.8	0.95	0.22	0.95	0.22
–2.4	0.91	0.29	0.91	0.29
–2	0.93	0.26	0.93	0.26
–1.6	0.92	0.28	0.92	0.27
–1.2	0.94	0.24	0.94	0.24
–0.8	0.91	0.29	0.91	0.29
–0.4	0.91	0.28	0.92	0.28
0	0.89	0.31	0.89	0.31
0.4	0.86	0.35	0.92	0.28
0.8	0.81	0.4	1	0
1.2	0.71	0.45	NA	NA
1.6	0.6	0.49	NA	NA
2	0.44	0.5	NA	NA
2.4	0.31	0.46	NA	NA
2.8	0.16	0.37	NA	NA
Total	0.75	0.43	0.92	0.27

Table A.23. Mean and SD for Classification Accuracy Conditional on θ_{PP} by Administration Method: All $\widehat{\boldsymbol{\theta}}_{\boldsymbol{RC}}$ vs $\widehat{\boldsymbol{\theta}}_{\boldsymbol{RC}} < \mathbf{0}$

	All $\hat{\theta}_{RC}$		$\hat{\theta}_{RC} < 0$	
θ_{PP}	Mean	SD	Mean	SD
–2.8	0.88	0.32	1	0
–2.4	0.87	0.34	1	0
–2	0.88	0.32	1	0
–1.6	0.84	0.37	1	0
–1.2	0.79	0.41	1	0
–0.8	0.73	0.45	0.97	0.17
–0.4	0.52	0.5	0.77	0.42
0	NA	NA	NA	NA
0.4	0.57	0.49	0.82	0.38
0.8	0.75	0.43	0.98	0.14
1.2	0.74	0.44	0.99	0.09
1.6	0.85	0.36	1	0
2	0.85	0.36	1	0
2.4	0.84	0.36	1	0
2.8	0.88	0.32	1	0
Total	0.75	0.43	0.92	0.27

Table A.24. Mean and SD for Test Length Conditional on θ_{RC} by Administration Method: All $\widehat{\boldsymbol{\theta}}_{\boldsymbol{RC}}$ vs $\widehat{\boldsymbol{\theta}}_{\boldsymbol{RC}} < \mathbf{0}$

	All $\hat{\theta}_{RC}$		$\hat{\theta}_{RC} < 0$	
θ_{RC}	Mean	SD	Mean	SD
–2.8	1.79	2.72	1.79	2.72
–2.4	2.11	3.65	2.11	3.66
–2	2.07	3.53	2.08	3.54
–1.6	2.73	4.92	2.68	4.78
–1.2	3.1	5.18	3.09	5.18
–0.8	4.45	7.26	4.41	7.2
–0.4	4.81	7.31	4.72	7.28
0	6.4	8.66	6.17	8.75
0.4	7.35	8.99	7.63	9.02
0.8	7.92	9.24	8	10.3
1.2	8.02	8.32	NA	NA
1.6	8.2	7.09	NA	NA
2	9.7	6.51	NA	NA
2.4	11.3	5.79	NA	NA
2.8	13	4.65	NA	NA
Total	6.2	7.41	3.28	5.76

Table A.25. Mean and SD for Classification Accuracy Conditional on θ_{PP} by Administration Method: All $\widehat{\boldsymbol{\theta}}_{\boldsymbol{RC}}$ vs $\widehat{\boldsymbol{\theta}}_{\boldsymbol{RC}} < \mathbf{0}$

	All $\widehat{\theta}_{RC}$		$\widehat{\theta}_{RC} < 0$	
θ_{PP}	Mean	SD	Mean	SD
–2.8	3.26	4.86	1.1	0.63
–2.4	3.56	5.16	1.13	0.8
–2	3.95	5.31	1.3	1.46
–1.6	4.58	5.99	1.46	1.64
–1.2	5.81	6.73	2.21	3.33
–0.8	7.53	7.87	3.46	5.49
–0.4	11.2	9.21	8.52	9.92
0	12.3	9.27	10	9.61
0.4	10.3	8.8	7.01	8.25
0.8	7.56	7.79	4.23	6.28
1.2	6.59	6.97	2.94	4.59
1.6	4.66	5.86	1.94	2.54
2	4.26	5.52	1.3	1.04
2.4	4.04	5.72	1.17	1.11
2.8	3.45	4.94	1.21	1.11
Total	6.2	7.41	3.28	5.76

Table A.26. Mean and SD for Classification Accuracy Conditional on θ_{RC} by Item Selection Method During Phase 2

	Fisher Information		Weighted Fisher Information	
θ_{RC}	Mean	SD	Mean	SD
–2.8	0.95	0.21	0.95	0.22
–2.4	0.92	0.27	0.91	0.29
–2	0.93	0.26	0.93	0.26
–1.6	0.92	0.27	0.92	0.27
–1.2	0.94	0.25	0.94	0.24
–0.8	0.91	0.28	0.91	0.29
–0.4	0.91	0.29	0.92	0.28
0	0.86	0.34	0.89	0.31
0.4	0.75	0.44	0.92	0.28
0.8	0.71	0.49	1	0
Total	0.92	0.27	0.92	0.27

Table A.27. Mean and SD for Classification Accuracy Conditional on θ_{PP} by Item Selection Method during Phase 2

	Fisher Information		Weighted Fisher Information	
θ_{PP}	Mean	SD	Mean	SD
–2.8	1	0	1	0
–2.4	1	0	1	0
–2	1	0.06	1	0
–1.6	1	0	1	0
–1.2	1	0.06	1	0
–0.8	0.96	0.2	0.97	0.17
–0.4	0.81	0.39	0.77	0.42
0	NA	NA	NA	NA
0.4	0.76	0.43	0.82	0.38
0.8	0.97	0.16	0.98	0.14
1.2	1	0.06	0.99	0.09
1.6	1	0.06	1	0
2	1	0	1	0
2.4	1	0	1	0
2.8	1	0	1	0
Total	0.92	0.27	0.92	0.27

Table A.28. Mean and SD for Test Length Conditional on θ_{RC} by Item Selection Method during Phase 2

	Fisher Information		Weighted Fisher Information	
θ_{RC}	Mean	SD	Mean	SD
–2.8	1.77	2.53	1.79	2.72
–2.4	2.07	3.49	2.11	3.66
–2	2.08	3.52	2.08	3.54
–1.6	2.59	4.48	2.68	4.78
–1.2	3.05	5.02	3.09	5.18
–0.8	4.89	7.87	4.41	7.2
–0.4	5.91	8.51	4.72	7.28
0	7.03	9.5	6.17	8.75
0.4	12.4	12.1	7.63	9.02
0.8	13	12.8	8	10.3
Total	3.61	6.35	3.28	5.76

Table A.29. Mean and SD for Test Length Conditional on θ_{PP} by Item Selection Method during Phase 2

	Fisher Information		Weighted Fisher Information	
θ_{PP}	Mean	SD	Mean	SD
–2.8	1.16	1	1.1	0.63
–2.4	1.17	0.84	1.13	0.8
–2	1.64	3.29	1.3	1.46
–1.6	1.65	2.3	1.46	1.64
–1.2	2.28	3.63	2.21	3.33
–0.8	3.85	6.39	3.46	5.49
–0.4	8.66	9.82	8.52	9.92
0	10.9	10.6	10	9.61
0.4	8.13	9.58	7.01	8.25
0.8	4.72	6.86	4.23	6.28
1.2	3.37	4.87	2.94	4.59
1.6	2.12	3.13	1.94	2.54
2	1.65	2.22	1.3	1.04
2.4	1.24	1.27	1.17	1.11
2.8	1.3	1.7	1.21	1.11
Total	3.61	6.35	3.28	5.76

Table A.30. Mean and SD for Classification Accuracy Conditional on θ_{RC} by Classification Stopping Rule for Phase 2

	CI		GLR		SPRT	
θ_{RC}	Mean	SD	Mean	SD	Mean	SD
–2.8	0.87	0.34	0.95	0.22	0.95	0.22
–2.4	0.84	0.37	0.91	0.29	0.91	0.28
–2	0.88	0.33	0.93	0.26	0.93	0.26
–1.6	0.87	0.34	0.92	0.27	0.92	0.27
–1.2	0.84	0.37	0.94	0.24	0.94	0.24
–0.8	0.82	0.38	0.91	0.29	0.9	0.29
–0.4	0.85	0.36	0.92	0.28	0.92	0.28
0	0.8	0.4	0.89	0.31	0.9	0.3
0.4	0.73	0.45	0.92	0.28	0.92	0.28
0.8	0.29	0.49	1	0	1	0
Total	0.85	0.36	0.92	0.27	0.92	0.27

Table A.31. Mean and SD for Classification Accuracy Conditional on θ_{PP} by Classification Stopping Rule during Phase 2

	CI		GLR		SPRT	
θ_{PP}	Mean	SD	Mean	SD	Mean	SD
–2.8	1	0	1	0	1	0
–2.4	1	0	1	0	1	0
–2	1	0	1	0	1	0
–1.6	1	0	1	0	1	0
–1.2	0.99	0.11	1	0	1	0
–0.8	0.89	0.32	0.97	0.17	0.97	0.17
–0.4	0.53	0.5	0.77	0.42	0.78	0.41
0	NA	NA	NA	NA	NA	NA
0.4	0.4	0.49	0.82	0.38	0.82	0.38
0.8	0.85	0.36	0.98	0.14	0.98	0.15
1.2	0.94	0.24	0.99	0.09	0.99	0.09
1.6	0.99	0.09	1	0	1	0
2	1	0	1	0	1	0
2.4	1	0	1	0	1	0
2.8	1	0	1	0	1	0
Total	0.85	0.36	0.92	0.27	0.92	0.27

Table A.32. Mean and SD for Test Length Conditional on θ_{RC} by Classification Stopping Rule for Phase 2

	CI		GLR		SPRT	
θ_{RC}	Mean	SD	Mean	SD	Mean	SD
–2.8	3.86	5.45	1.79	2.72	1.82	2.75
–2.4	4.61	6.64	2.11	3.66	2.22	3.8
–2	3.88	5.48	2.08	3.54	2.14	3.61
–1.6	5.23	7.2	2.68	4.78	2.75	4.88
–1.2	7.5	9.05	3.09	5.18	3.33	5.31
–0.8	9.87	10.2	4.41	7.2	5.04	7.39
–0.4	10.7	9.86	4.72	7.28	5.86	7.26
0	12.9	10.2	6.17	8.75	7.75	8.7
0.4	16.4	9.91	7.63	9.02	9.59	8.46
0.8	23.9	10.6	8	10.3	14.1	9.89
Total	7.13	8.76	3.28	5.76	3.73	5.97

Table A.33. Mean and SD for Test Length Conditional on θ_{PP} by Classification Stopping Rule for Phase 2

	CI		GLR		SPRT	
θ_{PP}	Mean	SD	Mean	SD	Mean	SD
–2.8	1.62	1.64	1.1	0.63	1.22	0.91
–2.4	1.83	2	1.13	0.8	1.35	1.29
–2	2.05	2.76	1.3	1.46	1.59	1.98
–1.6	2.41	3.08	1.46	1.64	1.77	2.21
–1.2	3.97	5.73	2.21	3.33	2.44	3.41
–0.8	7.19	8.92	3.46	5.49	3.89	5.86
–0.4	15.3	11.3	8.52	9.92	9.15	9.97
0	19.4	9.27	10	9.61	10.6	9.61
0.4	16.8	9.9	7.01	8.25	7.55	8.41
0.8	10.5	9.04	4.23	6.28	4.93	6.5
1.2	7.37	7.81	2.94	4.59	3.55	4.95
1.6	5.69	5.94	1.94	2.54	2.52	3.05
2	4.2	4.49	1.3	1.04	1.84	1.81
2.4	4.22	4.39	1.17	1.11	1.65	1.66
2.8	4.01	3.94	1.21	1.11	1.57	1.63
Total	7.13	8.76	3.28	5.76	3.73	5.97

Table A.34. Mean and SD for Classification Accuracy Conditional on θ_{RC} by Width of Indifference Region for Phase 2

	0.25		0.5		1	
θ_{RC}	Mean	SD	Mean	SD	Mean	SD
–2.8	0.92	0.28	0.95	0.22	0.95	0.21
–2.4	0.88	0.32	0.91	0.29	0.92	0.27
–2	0.9	0.3	0.93	0.26	0.93	0.25
–1.6	0.9	0.3	0.92	0.27	0.92	0.28
–1.2	0.9	0.3	0.94	0.24	0.94	0.23
–0.8	0.86	0.35	0.91	0.29	0.92	0.27
–0.4	0.88	0.32	0.92	0.28	0.92	0.27
0	0.83	0.37	0.89	0.31	0.93	0.26
0.4	0.81	0.39	0.92	0.28	0.95	0.22
0.8	0.57	0.53	1	0	1	0
Total	0.88	0.32	0.92	0.27	0.93	0.26

Table A.35. Mean and SD for Classification Accuracy Conditional on θ_{PP} by Width of Indifference Region for Phase 2

	0.25		0.5		1	
θ_{PP}	Mean	SD	Mean	SD	Mean	SD
–2.8	1	0	1	0	1	0
–2.4	1	0	1	0	1	0.06
–2	1	0	1	0	1	0.06
–1.6	1	0	1	0	1	0
–1.2	0.99	0.09	1	0	0.99	0.11
–0.8	0.92	0.27	0.97	0.17	0.96	0.2
–0.4	0.64	0.48	0.77	0.42	0.81	0.39
0	NA	NA	NA	NA	NA	NA
0.4	0.65	0.48	0.82	0.38	0.85	0.36
0.8	0.94	0.24	0.98	0.14	1	0.06
1.2	0.98	0.15	0.99	0.09	0.99	0.11
1.6	1	0	1	0	1	0.06
2	1	0	1	0	1	0
2.4	1	0	1	0	1	0
2.8	1	0	1	0	1	0
Total	0.88	0.32	0.92	0.27	0.93	0.26

Table A.36. Mean and SD for Test Length Conditional on θ_{RC} by Width of Indifference Region for Phase 2

	0.25		0.5		1	
θ_{RC}	Mean	SD	Mean	SD	Mean	SD
–2.8	2.9	4.58	1.79	2.72	1.18	0.96
–2.4	3.43	5.71	2.11	3.66	1.32	1.47
–2	2.85	4.65	2.08	3.54	1.28	1.35
–1.6	3.91	6.49	2.68	4.78	1.43	1.74
–1.2	5.06	7.7	3.09	5.18	1.61	2.1
–0.8	6.89	9.89	4.41	7.2	2.29	3.91
–0.4	6.89	9.41	4.72	7.28	2.47	4.05
0	8.71	10.8	6.17	8.75	3.1	4.95
0.4	10.5	11	7.63	9.02	4.12	5.59
0.8	15.1	14.2	8	10.3	4.71	6.85
Total	4.95	7.85	3.28	5.76	1.8	2.9

Table A.37. Mean and SD for Test Length Conditional on θ_{PP} by Width of Indifference Region for Phase 2

	0.25		0.5		1	
θ_{PP}	Mean	SD	Mean	SD	Mean	SD
–2.8	1.13	0.84	1.1	0.63	1.06	0.52
–2.4	1.27	1.13	1.13	0.8	1.04	0.31
–2	1.61	2.35	1.3	1.46	1.14	0.91
–1.6	1.91	2.47	1.46	1.64	1.23	1.01
–1.2	3.03	4.81	2.21	3.33	1.43	1.68
–0.8	5.62	8.09	3.46	5.49	2	3.07
–0.4	13	11.4	8.52	9.92	3.81	5.7
0	16.2	10.6	10	9.61	3.85	5.23
0.4	12.3	10.4	7.01	8.25	3.03	4.41
0.8	6.4	7.97	4.23	6.28	2.13	3.39
1.2	4.53	6.55	2.94	4.59	1.66	2.18
1.6	2.62	3.81	1.94	2.54	1.25	1.07
2	1.66	1.89	1.3	1.04	1.08	0.45
2.4	1.4	1.55	1.17	1.11	1.07	0.48
2.8	1.36	1.66	1.21	1.11	1.11	0.83
Total	4.95	7.85	3.28	5.76	1.8	2.9

Table A.38. Mean and SD for Misclassification Rate Conditional on θ_{RC} by Width of Indifference Region for Phase 2

	0.25		0.5		1	
θ_{RC}	Mean	SD	Mean	SD	Mean	SD
–2.8	0.01	0.12	0.03	0.18	0.05	0.21
–2.4	0.02	0.15	0.07	0.25	0.08	0.27
–2	0.02	0.15	0.04	0.19	0.07	0.25
–1.6	0.02	0.15	0.05	0.23	0.08	0.28
–1.2	0.02	0.14	0.04	0.21	0.06	0.23
–0.8	0.02	0.14	0.06	0.24	0.07	0.26
–0.4	0.03	0.17	0.05	0.22	0.08	0.27
0	0.01	0.1	0.04	0.2	0.07	0.26
0.4	0.02	0.13	0.02	0.13	0.03	0.18
0.8	0	0	0	0	0	0
Total	0.02	0.14	0.05	0.21	0.07	0.25

Table A.39. Mean and SD for Misclassification Rate Conditional on θ_{PP} by Width of Indifference Region for Phase 2

	0.25		0.5		1	
θ_{PP}	Mean	SD	Mean	SD	Mean	SD
–2.8	0	0	0	0	0	0
–2.4	0	0	0	0	0	0.06
–2	0	0	0	0	0	0.06
–1.6	0	0	0	0	0	0
–1.2	0	0	0	0	0.01	0.11
–0.8	0.01	0.09	0.02	0.14	0.04	0.2
–0.4	0.02	0.14	0.09	0.29	0.17	0.38
0	NA	NA	NA	NA	NA	NA
0.4	0.03	0.18	0.1	0.3	0.15	0.35
0.8	0	0	0	0	0	0.06
1.2	0	0	0	0.06	0.01	0.11
1.6	0	0	0	0	0	0.06
2	0	0	0	0	0	0
2.4	0	0	0	0	0	0
2.8	0	0	0	0	0	0
Total	0.02	0.14	0.05	0.21	0.07	0.25

Table A.40. Mean and SD for Inconclusive Rate Conditional on θ_{RC} by Width of Indifference Region for Phase 2

	0.25		0.5		1	
θ_{RC}	Mean	SD	Mean	SD	Mean	SD
–2.8	0.07	0.26	0.02	0.13	0	0
–2.4	0.1	0.3	0.02	0.15	0	0
–2	0.08	0.27	0.03	0.18	0	0
–1.6	0.08	0.27	0.03	0.16	0	0
–1.2	0.08	0.27	0.02	0.13	0	0.04
–0.8	0.12	0.33	0.03	0.18	0	0.06
–0.4	0.09	0.28	0.03	0.18	0	0.05
0	0.16	0.36	0.06	0.25	0	0.06
0.4	0.17	0.38	0.07	0.25	0.02	0.13
0.8	0.43	0.53	0	0	0	0
Total	0.09	0.29	0.03	0.17	0	0.04

Table A.41. Mean and SD for Inconclusive Rate Conditional on θ_{PP} by Width of Indifference Region for Phase 2

	0.25		0.5		1	
θ_{PP}	Mean	SD	Mean	SD	Mean	SD
–2.8	0	0	0	0	0	0
–2.4	0	0	0	0	0	0
–2	0	0	0	0	0	0
–1.6	0	0	0	0	0	0
–1.2	0.01	0.09	0	0	0	0
–0.8	0.07	0.26	0.01	0.11	0	0
–0.4	0.34	0.48	0.14	0.34	0.02	0.13
0	NA	NA	NA	NA	NA	NA
0.4	0.32	0.47	0.08	0.27	0	0.07
0.8	0.06	0.24	0.02	0.14	0	0
1.2	0.02	0.15	0	0.06	0	0
1.6	0	0	0	0	0	0
2	0	0	0	0	0	0
2.4	0	0	0	0	0	0
2.8	0	0	0	0	0	0
Total	0.09	0.29	0.03	0.17	0	0.04

Table A.42. Mean and SD for Classification Accuracy Conditional on θ_{RC} by Maximum Test Length for Phase 2

	25/40		25+15	
θ_{RC}	Mean	SD	Mean	SD
–2.8	0.95	0.22	0.95	0.23
–2.4	0.91	0.29	0.91	0.29
–2	0.93	0.26	0.93	0.26
–1.6	0.92	0.27	0.91	0.28
–1.2	0.94	0.24	0.91	0.29
–0.8	0.91	0.29	0.86	0.35
–0.4	0.92	0.28	0.87	0.34
0	0.89	0.31	0.83	0.37
0.4	0.92	0.28	0.81	0.39
0.8	1	0	0.71	0.49
Total	0.92	0.27	0.9	0.3

Table A.43. Mean and SD for Classification Accuracy Conditional on θ_{PP} by Maximum Test Length for Phase 2

	25/40		25+15	
θ_{PP}	Mean	SD	Mean	SD
–2.8	1	0	1	0
–2.4	1	0	1	0
–2	1	0	1	0.06
–1.6	1	0	1	0.06
–1.2	1	0	0.98	0.13
–0.8	0.97	0.17	0.93	0.26
–0.4	0.77	0.42	0.68	0.47
0	NA	NA	NA	NA
0.4	0.82	0.38	0.72	0.45
0.8	0.98	0.14	0.93	0.26
1.2	0.99	0.09	0.97	0.18
1.6	1	0	0.99	0.09
2	1	0	1	0
2.4	1	0	1	0
2.8	1	0	1	0
Total	0.92	0.27	0.9	0.3

Table A.44. Mean and SD for Test Length Conditional on θ_{RC} by Maximum Test Length for Phase 2

	25/40		25+15	
θ_{RC}	Mean	SD	Mean	SD
–2.8	1.79	2.72	1.78	2.67
–2.4	2.11	3.66	2.01	3.06
–2	2.08	3.54	2.01	3.16
–1.6	2.68	4.78	2.43	3.69
–1.2	3.09	5.18	2.75	3.88
–0.8	4.41	7.2	3.5	4.59
–0.4	4.72	7.28	3.83	4.74
0	6.17	8.75	4.7	5.37
0.4	7.63	9.02	5.81	5.37
0.8	8	10.3	6	6.35
Total	3.28	5.76	2.81	4.03

Table A.45. Mean and SD for Test Length Conditional on θ_{PP} by Maximum Test Length for Phase 2

	25/40		25+15	
θ_{PP}	Mean	SD	Mean	SD
–2.8	1.1	0.63	1.1	0.63
–2.4	1.13	0.8	1.13	0.8
–2	1.3	1.46	1.3	1.42
–1.6	1.46	1.64	1.45	1.53
–1.2	2.21	3.33	2.13	2.95
–0.8	3.46	5.49	3.09	4.14
–0.4	8.52	9.92	6.33	6.02
0	10	9.61	7.74	6.09
0.4	7.01	8.25	5.74	5.59
0.8	4.23	6.28	3.68	4.56
1.2	2.94	4.59	2.68	3.44
1.6	1.94	2.54	1.94	2.5
2	1.3	1.04	1.3	1.04
2.4	1.17	1.11	1.17	1.11
2.8	1.21	1.11	1.21	1.11
Total	3.28	5.76	2.81	4.03

Appendix B: Supplementary Figures

Figure B.1. Joint Distributions of True θ Values and Estimated θ Values, with GRM as the Model for $\widehat{\theta}_{PP}$ vs. 2PL as the Model for $\widehat{\theta}_{PP}$

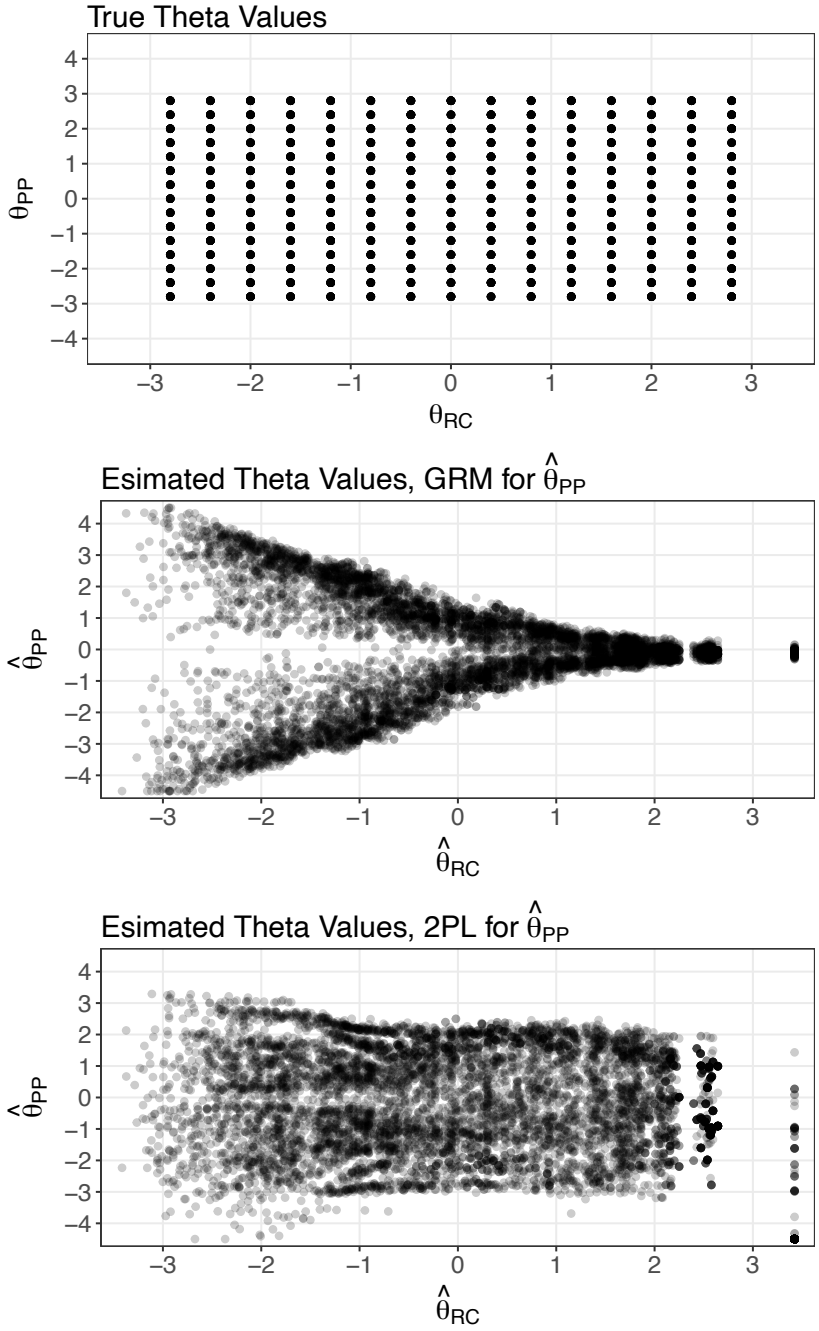


Figure B.2. Misclassification Rate and Inconclusive Classification Rate as a Function of Including vs. Deleting Cases with $\widehat{\theta}_{RC} < 0$ Conditional on θ_{RC} (top) and θ_{PP} (bottom)

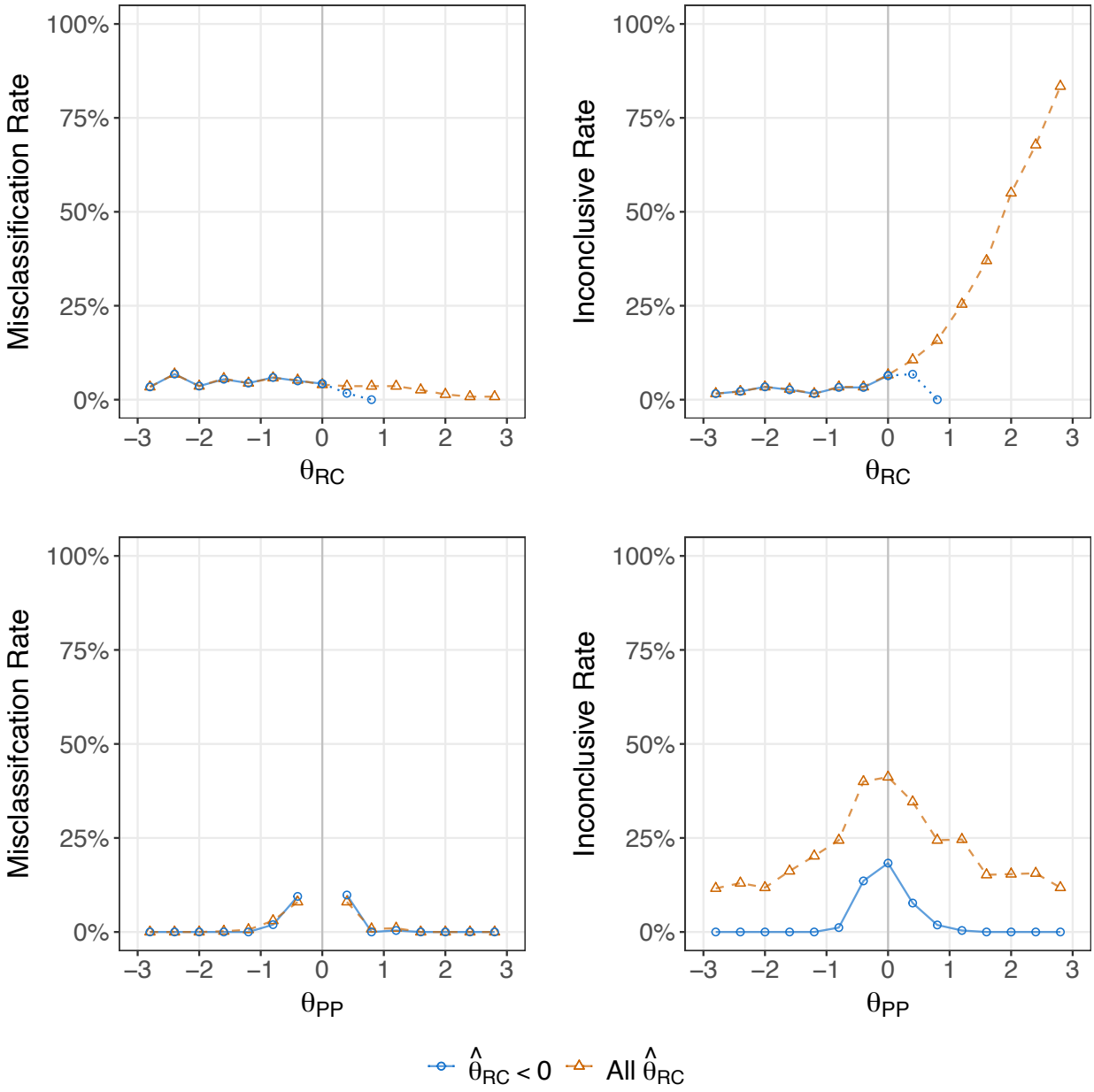


Figure B.3. Misclassification Rate and Inconclusive Classification Rate as a Function of Fisher Information vs. Weighted Fisher Information Conditional on θ_{RC} (top) and θ_{PP} (bottom)

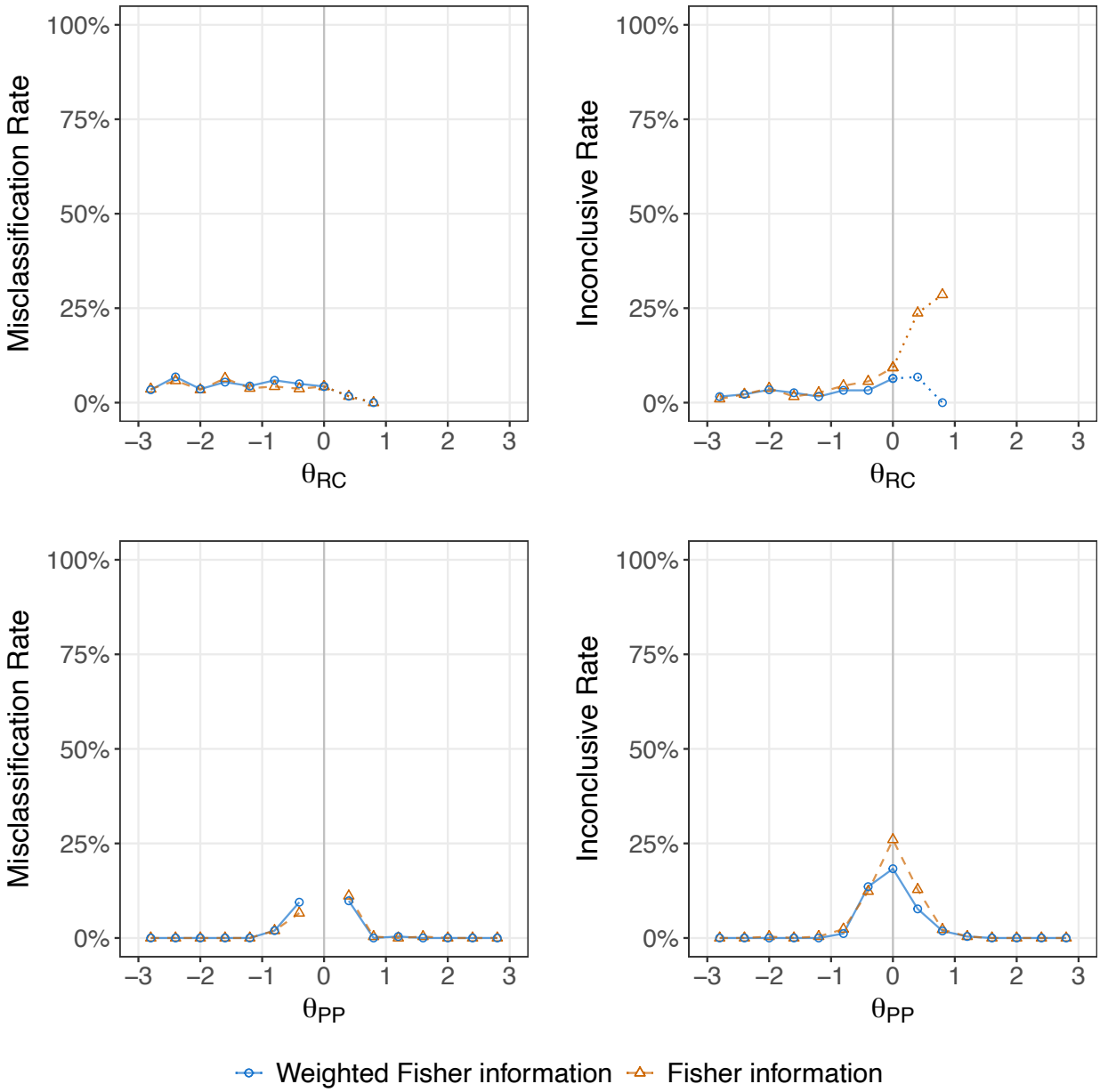


Figure B.4. Misclassification Rate and Inconclusive Classification Rate as a Function of Classification Stopping Rule Conditional on θ_{RC} (top) and θ_{PP} (bottom)

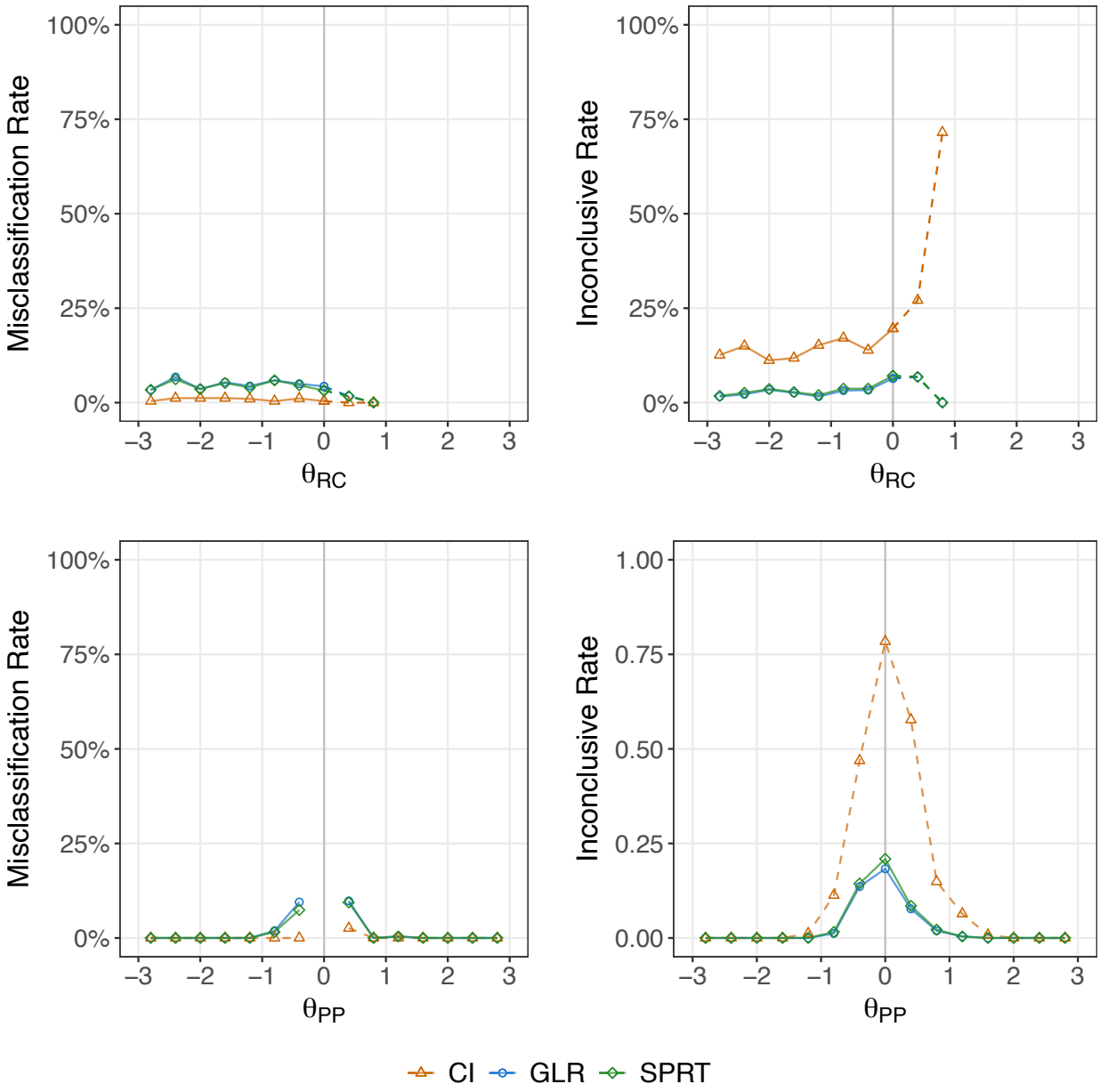


Figure B.5. Misclassification Rate and Inconclusive Classification Rate as a Function of Maximum Phase 2 Test Length Conditional on θ_{RC} (top) and θ_{PP} (bottom)

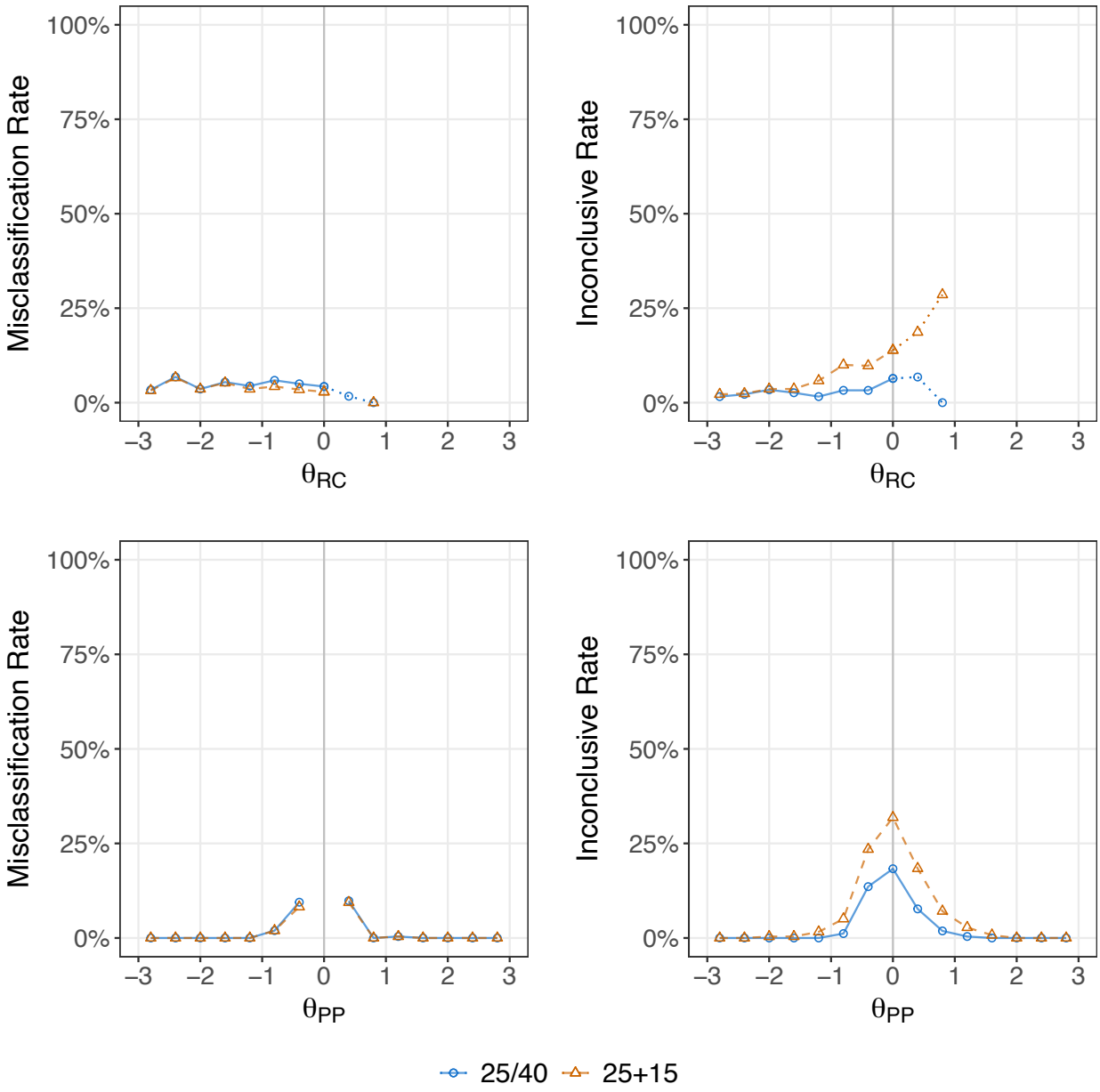


Figure B.6. Average Total MOCCA Test Length as a Function of Including vs. Deleting Cases with $\widehat{\theta}_{RC} < 0$ Conditional on θ_{RC} and θ_{PP}

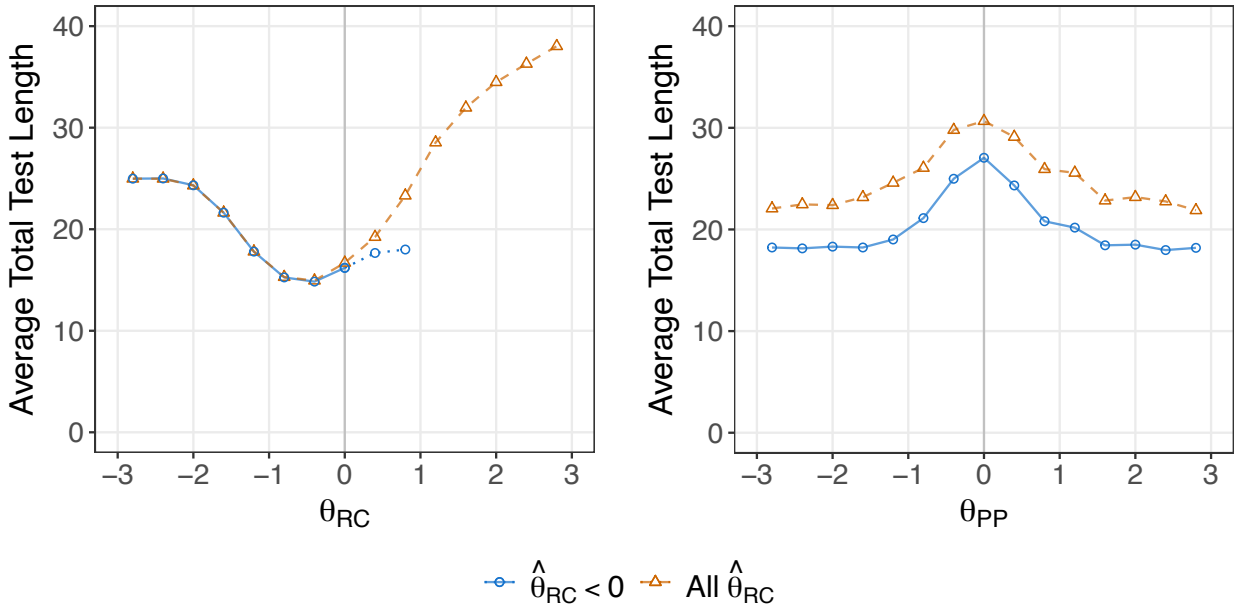


Figure B.7. Average Total MOCCA Test Length as a Function of Fisher Information vs. Weighted Fisher Information Conditional on θ_{RC} and θ_{PP}

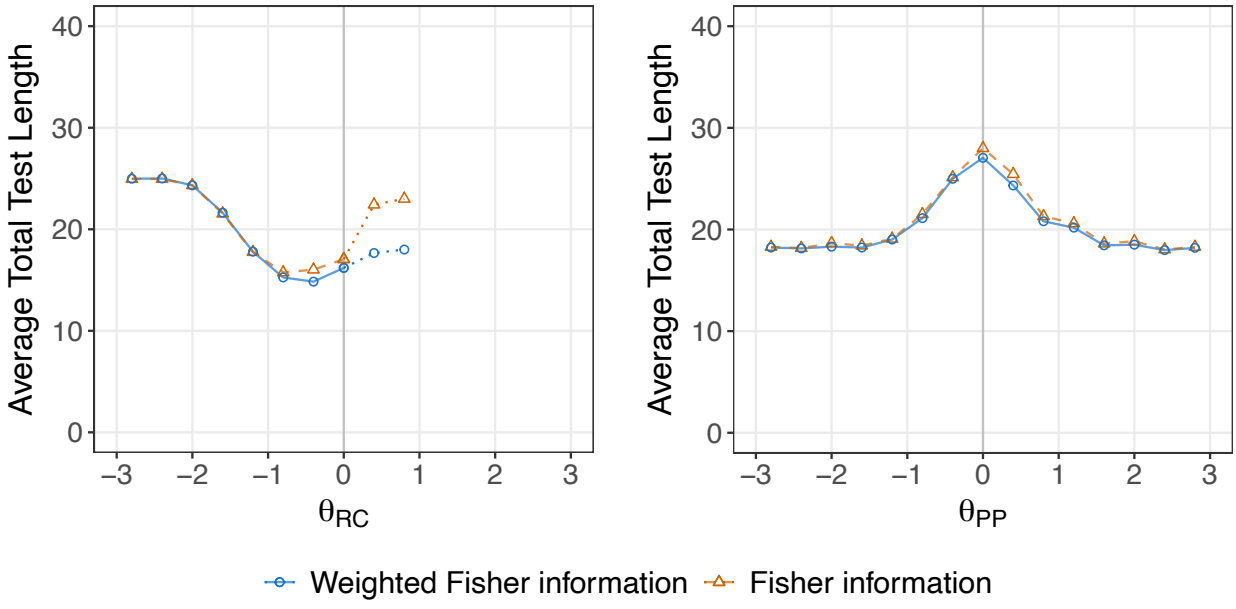


Figure B.8. Average Total MOCCA Test Length as a Function of Classification Stopping Rule Conditional on θ_{RC} and θ_{PP}

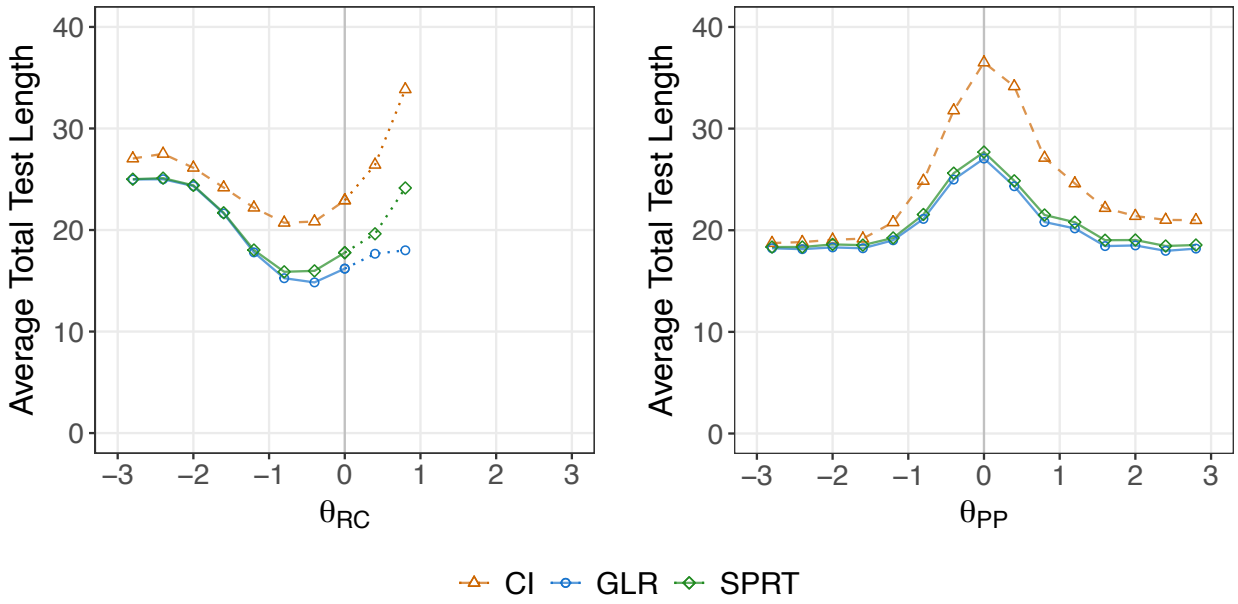


Figure B.9. Average Total MOCCA Test Length as a Function of GLR Indifference Region Conditional on θ_{RC} and θ_{PP}

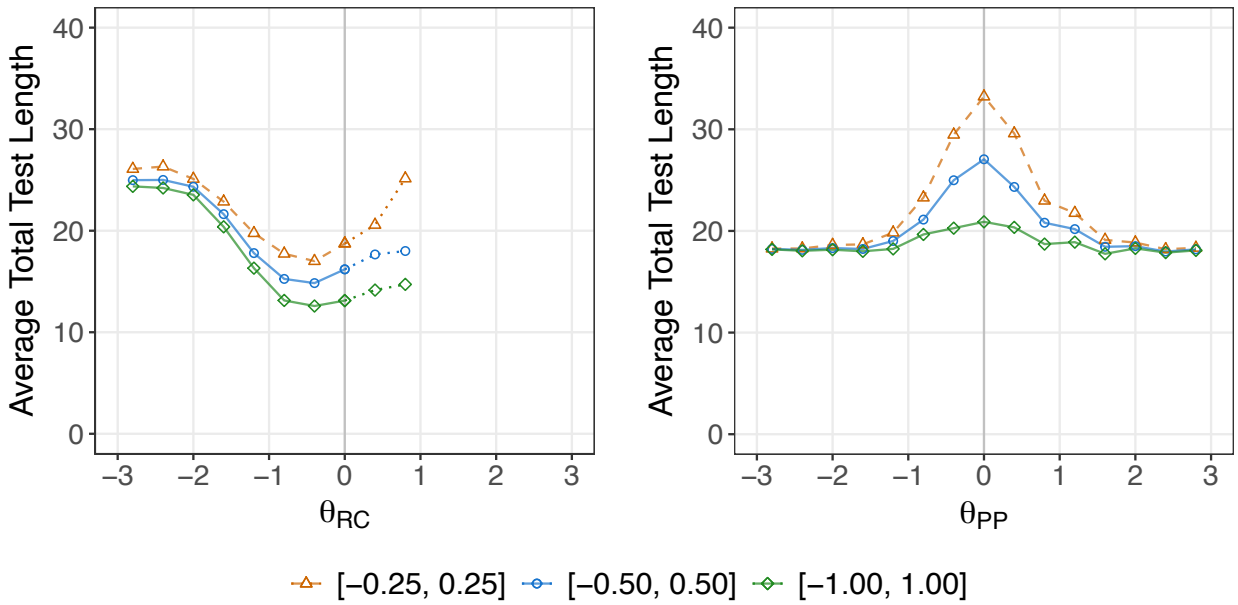
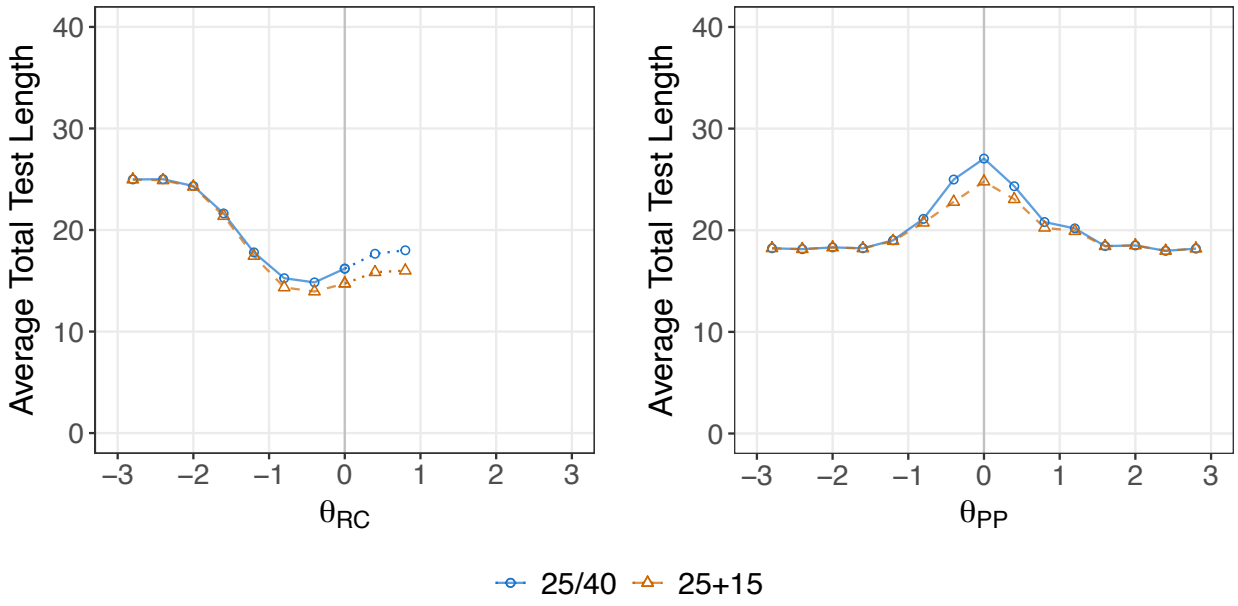


Figure B.10. Average Total MOCCA Test Length as a Function of Maximum Phase 2 Test Length Conditional on θ_{RC} and θ_{PP}



Appendix C: ANOVA Model

Mixed Design ANOVA Model

All ANOVAs were run as mixed design ANOVAs with the following form of linear model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \tau_{k(i)} + \varepsilon_{ijk}$$

where:

Y_{ijk} is the dependent variable for simulee *k* with between-group variable level *i* and within-group variable level *j*,

μ is the grand mean,

α_i is the effect of between-group variable level *i*,

β_j is the effect of within-group variable level *j*,

$(\alpha\beta)_{ij}$ is the interaction effect of between-group variable level *i* and within-group variable level *j*,

$\tau_{k(i)}$ is the between-group error, and

ε_{ijk} is the within-group error.

For Phase 1 analyses, the set of dependent variables is {Bias, RMSE, RMS-SEM, Average Test Length), the set of between-group variables is $\{\theta_{RC}\}$, and the set of within-group variables is {Stopping Rule, Maximum Test Length, Number of Response Options}. For Phase 2 analyses, the set of dependent variables is {Classification Accuracy, Average Test Length}, the set of between-group variables is $\{\theta_{RC}, \theta_{PP}\}$, and the set of within-group variables is {Item Selection Rule, Stopping Rule, Indifference Region, Upper Limit on Number of Items}.

Calculation of ω^2 for Mixed Design ANOVA

$$\omega^2_{between} = \frac{df_{between\ factor} * \left(MS_{between\ factor} - MS_{between\ error} \right)}{SS_{total} + MS_{between\ error}}$$

$$\omega^2_{within} = \frac{df_{within\ factor} * \left(MS_{within\ factor} - MS_{within\ error} \right)}{SS_{total} + MS_{between\ error}}$$

$$\omega^2_{interaction} = \frac{df_{between*within\ factor} * \left(MS_{between\ factor} - MS_{within\ error} \right)}{SS_{total} + MS_{between\ factor}}$$

where sums of squares total (SS_{total}) is defined as

$$SS_{total} = \left(df_{between\ factor} * MS_{between\ factor}\right) + \left(df_{within\ factor} * MS_{within\ factor}\right) + \left(df_{between*within\ factor} * MS_{between*within\ factor}\right) + \left(df_{between\ error} * MS_{between\ error}\right) + \left(df_{within\ error} * MS_{within\ error}\right)$$