Article

Journal of Educational and Behavioral Statistics
2023, Vol. 48, No. 6, pp. 914–941
DOI: 10.3102/10769986231158301
Article reuse guidelines: sagepub.com/journals-permissions
© 2023 AERA. https://journals.sagepub.com/home/jeb

A Diagnostic Tree Model for Adaptive Assessment of Complex Cognitive Processes Using Multidimensional Response Options

Mark L. Davison
David J. Weiss
University of Minnesota

Joseph N. DeWeese
University of Minnesota

Ozge Ersan
Turkish Ministry of National Education

Gina Biancarosa
Patrick C. Kennedy
University of Oregon

A tree model for diagnostic educational testing is described along with Monte Carlo simulations designed to evaluate measurement accuracy based on the model. The model is implemented in an assessment of inferential reading comprehension, the Multiple-Choice Online Causal Comprehension Assessment (MOCCA), through a sequential, multidimensional, computerized adaptive testing (CAT) strategy. Assessment of the first dimension, reading comprehension (RC), is based on the three-parameter logistic model. For diagnostic and intervention purposes, the second dimension, called process propensity (PP), is used to classify struggling students based on their pattern of incorrect responses. In the simulation studies, CAT item selection rules and stopping rules were varied to evaluate their effect on measurement accuracy along dimension RC and classification accuracy along dimension PP. For dimension RC, methods that improved accuracy tended to increase test length. For dimension PP, however, item selection and stopping rules increased classification accuracy without materially increasing test length. A small live-testing pilot study confirmed some of the findings of the simulation studies. Development of the assessment has been guided by psychometric theory, Monte Carlo simulation results, and a theory of instruction and diagnosis.

Keywords: *diagnostic testing; computerized adaptive testing; reading assessment; item response models; classification testing; formative assessment*

As stated in the call for papers, “The purpose of this issue is to highlight statistical methods for providing decision makers and users with fine-grained information to improve educational and behavioral outcomes.” The goal is to “advance methods that are consistent with an assessment framework of ‘diagnose and intervene’ rather than the paradigm of ‘rank and sort.’” To achieve these goals, an assessment framework must go well beyond a theory of latent variables embedded in a statistical model. It must include a theory of intervention and real items, both of which are linked to the latent variables in a statistical model. This article describes the development of a reading comprehension (RC) test, MOCCA. MOCCA adopts a diagnostic item response theory (IRT) approach for the adaptive measurement of molecular cognitive processes that are not readily decomposed into separate skills. It is based on intervention research, unique items and multiple-choice response types, and a statistical tree model of those responses. Employing a computerized adaptive testing (CAT) strategy, it provides overall inferential comprehension scores along a latent dimension and a diagnostic classification for struggling readers. This classification provides information to assist teachers in individualizing additional instruction for struggling readers. MOCCA is more than just an idealized, statistical method for the future. It is a method that has already led to a real, online assessment. For a demonstration, see https://blogs.uoregon.edu/mocca/.

Introduction

In the sections that follow, we first describe the intervention research on which the assessment is based. Next, we describe the assessment itself: its purposes, items, multiple-choice response structure, statistical model, and CAT administration strategy. Finally, we present the results from Monte Carlo simulations that guided the design of the new CAT administration strategy. Readers interested in more information about earlier versions of the assessment, their reliability, and their validity should consult Biancarosa et al. (2019), Carlson et al. (2014), Davison et al. (2019), Liu et al. (2019), and Su and Davison (2019).

Theory of Intervention

According to prior research, there are two types of struggling readers: those who struggle with prereading skills (e.g., phonemic awareness, word identification) and those who struggle with comprehension (Cain & Oakhill, 2006; Perfetti, 2007). Although readers who struggle specifically with comprehension, conservatively about 7% to 10% of all readers (Catts et al., 2012), may do so for a variety of reasons, one of the primary distinguishing characteristics of readers with specific poor comprehension is difficulty with generating inferences that establish global coherence (e.g., Currie & Cain, 2015; Pimperton & Nation, 2010; Spencer et al., 2019). That is, where a text requires readers to make a specific inference for coherent understanding, readers with specific poor comprehension tend not to make these inferences.

Think-aloud research has shown that readers with specific poor comprehension instead tend to rely on one of two cognitive processes during reading (e.g., Carlson et al., 2014; McMaster et al., 2012; Rapp et al., 2007). The first is a tendency to paraphrase information in the text rather than make an inference. The second is a tendency to make an elaborative inference based on information in the text or background knowledge which, although it may lead to a more enriched mental model of what is being read, does not establish global coherence. Whereas global coherence inferences are necessary for comprehension, these elaborative inferences are unnecessary. Moreover, McMaster et al. (2012) found in a randomized experiment that paraphrasing poor comprehenders benefited more than elaborators from a general questioning condition to encourage readers to make general connections during reading. In contrast, elaborating comprehenders benefited more than paraphrasers from a causal questioning strategy to encourage readers to make globally coherent inferences during reading. These results suggest the efficacy of focusing instruction for elaborating and paraphrasing poor comprehenders on questioning strategies matched to their cognitive process propensity (PP).

MOCCA

In response to the research described above, a new assessment called MOCCA was developed and validated as a means of providing teachers with information on their students' RC processes (Carlson et al., 2014; Davison et al., 2018). Specifically, MOCCA can be used (a) as a general outcome measure for inferential RC, (b) to track students' progress within and across the Grade 3 to 6 years, and (c) as a formative reading assessment that both identifies poor comprehenders and diagnoses their propensity to rely on a paraphrasing or elaborating approach. MOCCA has been revised into a CAT to facilitate and enhance its suitability for all three of these purposes by improving the precision of scores while decreasing test length.

MOCCA's diagnostic classifications can be used by teachers to individualize instruction for poor comprehenders. Specifically, MOCCA interpretive guidance provides teachers with different questioning strategies aligned to McMaster et al.'s (2014) approaches, depending on whether a student has a paraphrasing or elaborating propensity. Moreover, MOCCA flags students who do not have a clear propensity for either paraphrasing or elaborating, and the interpretive guide directs teachers to triangulate MOCCA results with data from other reading measures to identify potential problems in component reading skills (e.g., word reading, fluency, literal comprehension) and vocabulary. That is, MOCCA is designed to diagnose problems in comprehension. When students do not show clear diagnosable patterns, the assessment suggests further diagnostic assessment to determine alternative root causes of poor comprehension skills.

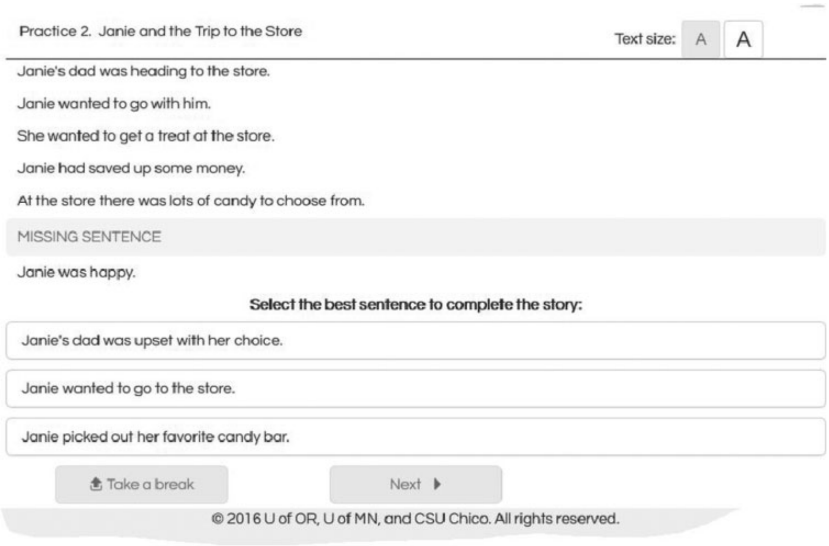
MOCCA Item Tasks and Responses

Each MOCCA item consists of a short story followed by three types of responses. The story contains between 5 and 10 sentences with the second to last sentence missing. From the three to five response alternatives, students must identify the sentence that best completes the story. Whereas most multiple-choice tests contain two types of answers (i.e., correct and incorrect), each MOCCA item has three types of responses: the correct response and two different types of incorrect. The correct answer (i.e., the causal coherent response) requires an inference from the information in the passage that leads to identification of the response sentence that best completes the passage. The first type of incorrect response (i.e., the paraphrase response) simply repeats or paraphrases information in the passage. It does not involve an inference, and it does not advance the story in a way that would complete the story. The second type of incorrect response (i.e., the elaboration response) does involve an inference that elaborates the story, but it does not complete the story line and may even somewhat contradict the story line. Each item has a correct response and either one paraphrase and one elaboration response or two of each incorrect response type.

Figure 1 shows a sample MOCCA item entitled “Janie and the Trip to the Store.” Note that 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 because it merely reiterates information explicitly stated earlier in the story. The third alternative is the correct, causal coherent, 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 identifies whether she accomplished her goal of getting a treat.

In addition to containing three types of response alternatives for every item, MOCCA items differ from those usually seen in RC tests in one other important respect. Many reading tests contain passages with several items related to each passage. Because the several items for a single passage all refer to the same passage, they form testlets that might violate the local independence assumption of IRT. In MOCCA, there is only one item for each story, so the structure of the item does not violate 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 CAT format, particularly compared to other reading tests.

According to the cognitive theory underlying MOCCA, respondents can apply one of the three processes to arrive at a response, one of which will result in the



Practice 2. Janie and the Trip to the Store

Janie's dad was heading to the store.

Janie wanted to go with him.

She wanted to get a treat at the store.

Janie had saved up some money.

At the store there was lots of candy to choose from.

MISSING SENTENCE

Janie was happy.

Select the best sentence to complete the story:

Janie's dad was upset with her choice.

Janie wanted to go to the store.

Janie picked out her favorite candy bar.

© 2016 U of OR, U of MN, and CSU Chico. All rights reserved.

FIGURE 1. *Sample MOCCA item.*

correct answer and two of which will result in an incorrect answer. The probability of choosing a particular response to an item corresponds to the probability of applying the corresponding cognitive process. As described in the following, the approach utilizes CAT and IRT-based classification to minimize testing time (Weiss, 1982). In agreement with the recent suggestion of Sireci (2022), this approach is less about grading and more about learning and instruction.

The MOCCA Model Versus Cognitive Diagnostic Models

Diagnostic testing often uses cognitive diagnostic test models (CDMs), such as the deterministic inputs, noisy “and” gate (DINA; Junker & Sitsma, 2001) model or the deterministic inputs, noisy “or” gate (DINO; Templin & Henson, 2006) model. In general, CDM models the probability of a correct answer as a function of latent categorical variables, usually associated with mastery or nonmastery of certain skills. The DINA model is an example of a noncompensatory model, meaning mastery of all skills related to the item are required to increase the probability of a correct response. In contrast, the DINO model is a compensatory model, where mastery of one skill can compensate for nonmastery of other skills.

The most familiar of these models has four features that distinguish it from the approach discussed here. First, the model assumes that the underlying cognitive process readily decomposes into several discrete subskills (Junker & Sijtsma, 2001). Second, most of the models (but not Chen & de la Torre, 2013 or von Davier, 2008) assume that skills are dichotomous; the student either does or does not possess each skill. CDMs coarsely classify respondents into two or a few categories, such as masters or nonmasters, relative to skill dimensions. Third, CDMs assume multidimensionality at the item level, rather than at the item response level, such that each item (rather than each response option) is associated with one or more of the dichotomous dimensions. Fourth, the CDM models describe students in terms of probabilities of belonging to response classes that are only indirectly related to the probability of subskill mastery or probability of correctly answering an item. Although there have been recent developments in CAT based on CDMs, research on actual implementation is in an early stage (Yu et al., 2019). Such models do not readily apply to complex cognitive processes that map onto response options, rather than items, and that are not readily decomposed into separate dichotomous (or polytomous) subskills. The CDM literature focuses on statistical models of responses to items rather than response option content.

A Focus on Response Options

In statistics and the sciences, researchers have focused on response option content as a way to make tests more about learning and less about grading (e.g., Delmas et al., 2007; Hermann-Abell & DeBoear, 2011; Hestenes et al., 1992; Sadler, 1998). In this literature, alternatives are written to represent common misconceptions, and the tests include scores that report the occurrences of various misconceptions in a student's responses. When distractors correspond to misconceptions, Hestenes et al. (1992) use the term "distractor-driven assessment" and Sadler uses the term "concept assessment." In practice, however, each misconception appears in only a small number of items. If a particular type of distractor seldom appears—say it appears only twice—evaluating the student's tendency to commit the misconception is a bit like judging the student's free-throw shooting ability based on two free-throw attempts. The information available is not very reliable.

Distractor-driven or concept assessment requires a rethinking of psychometric theory in terms of nesting. Response options are nested within items that are nested within a test. To complicate matters, response options constitute a set of mutually exclusive and exhaustive categories, so there is a linear dependence of any one option on all the others. Originally, psychometric theory focused on the test level: classical test theory. IRT focuses on the item level. Distractor-driven and concept assessments require more focus on the lowest level of the hierarchy, the response option.

At least two models have been proposed for modeling response options within items. When focusing on response options, each item has a response vector $\mathbf{x}_j$ (rather than a response variable x_j), and the overall model includes a model for each item response variable within the item vector $\mathbf{x}_j$. Johnson and Bolt (2010) described a response vector for each item with each variable in the vector corresponding to a response category along a polytomous item. They then proposed a highly parameterized multinomial multifactor model to account for individual differences in traits and response style. Their model includes the multidimensional random coefficients multinomial logit model of Adams et al. (1997) as a special case. Their model applies to a very different application (polytomous items) rather than the present model for multiple-choice items.

Bradshaw and Templin (2014) proposed a model for distractor-driven items with a continuous dimension for the composite ability and dichotomous dimensions corresponding to misconceptions, in which a misconception is present or absent. It is a hybrid of an item response model and a cognitive diagnostic model. It appears to require large sample sizes, as the simulation study included sample sizes of 3,000 and 10,000, and the real data example included a sample of 10,039. CAT based on the model remains a question for future research.

Early in our research on adapting MOCCA to a CAT approach, we researched nominal and ordered category unidimensional IRT models assuming three polytomous response categories: causal coherent, paraphrase, and elaboration. In research on the nominal model, two trends emerged suggesting an ordered category model. First, when we plotted empirical category response functions, the paraphrase function was monotonically decreasing, the elaboration function was a nonmonotone, single-peaked function, and the causal coherent was monotone increasing (see Figure 2 in Liu et al., 2019). Second, in the nominal model, for most items (but not all), the discrimination parameters for the categories were ordered paraphrase < elaboration < causal coherent. These two findings led us to reject the unidimensional nominal model in favor of a unidimensional ordered category model, at least if one limits comparisons to unidimensional models. We also compared fit measures for unidimensional models and a two-dimensional tree model (Davison et al., 2017). Both fit reasonably well and fit measures were not decisive. In thinking about dimensionality, we also studied whether incorrect answers added information over and above that provided by correct answers in the identification of students at-risk of not reaching proficiency on a statewide exam (Biancarosa et al., 2019). In this earlier research, we decided on a two-dimensional tree model partly based on fit measures but also because the tree model provided a way to isolate and report the extra information provided by incorrect answers in the second dimension, here called the PP dimension.

There are a number of possibly plausible CDM models for our data, including hybrid CDM/IRT (e.g., Hong et al., 2015) and polytomous (e.g., Chen & de la Torre, 2013; von Davier, 2008) models. However, they were designed as models for items, not item response options, and may not perform well with the large amount of missing data for some response options, as observed in response vectors of MOCCA items.

We thus elected to implement and evaluate a tree model (De Boeck et al., 2017; De Boeck & Partchev, 2012; Kim, 2022; Kim & Bolt, 2021; Partchev & De Boeck, 2012). By adding or subtracting branches, tree models can be adapted to a great many applications. De Boeck and colleagues (De Boeck et al., 2017; De Boeck & Partchev, 2012; Partchev & De Boeck, 2012) have used tree models to study item responses differing in their response times. However, a tree model can also be used to study differences among correct answer response types or differences among incorrect answer response types, as described in the following.

The Adaptive Diagnostic Tree Model

IRT trees are decision trees with a node for each decision and a latent dimension for each node that accounts for the probability of the decision at that node. Figure 2 shows the tree diagram for a MOCCA item with a node for the choice between the correct and incorrect alternatives and a second node for the choice between incorrect alternatives if the correct choice is not made. The model for each item posits two decision nodes with a dimension to account for the correct decision and a second dimension to account for the choice of incorrect response given an incorrect choice.

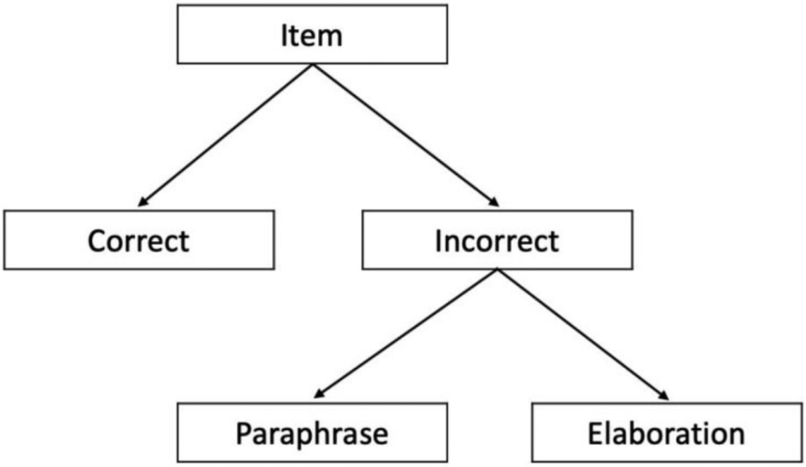


FIGURE 2. *MOCCA two-level tree model.*

IRT decision trees share a common missing data problem. For decisions beyond the first node, an item will provide information about that decision only if some condition is satisfied. For instance, for our second node, the item will provide information about a student's probability of choosing a paraphrase over an elaboration response only for items where the student selects an incorrect response. Because of this missingness, a person can be located along the second dimension with less precision. Given the loss of precision, a coarser level of feedback is often used regarding that second dimension. That is, rather than report a score, a classification is implemented. Adaptive classification (Wang et al., 2021) is used to classify each person into one of the three categories along the PP dimension based on their predominant incorrect response type (paraphrase, elaboration, and indeterminate). Here, the PP dimension is viewed as bipolar with students who only choose paraphrase incorrect responses at the positive end, students who only choose elaboration at the negative end, and students equally likely to choose either one in the middle ($\theta = 0$). That is, $\theta = 0$ is a cut-score that divides students with a probability greater than .5 of choosing paraphrase for an item of average difficulty from students with a probability greater than .5 of choosing elaboration for that same item. Given the missing data (correct responses) with respect to the PP dimension, it was appropriate to adopt a coarse classification, more commonly associated with CDMs, rather than a refined score.

The simulation studies reported in the following examined the accuracy of scores along the RC dimension, the accuracy of classification along the PP dimension, and a trade-off between the two in adaptively selecting items for the purposes of accurately locating the student along the RC dimension versus accurately classifying the student along the PP dimension. To date, we have prioritized study of person parameter estimation to the neglect of item calibration, a limit of our work thus far. The approach involved adaptive measurement along the RC dimension followed by adaptive classification along the PP dimension. A small live-testing pilot study was implemented to confirm some of the conclusions from the simulation studies.

Models

In tree models, an item response generates a small vector of response variables, not a single response variable. In the present case, there are two response variables for each item *j*, $\mathbf{X}_{ij} = (X_{1ij}, X_{2ij})$. In Phase 1 of the adaptive testing, items are selected because their first response variable X_{1ij} will maximize Fisher information along dimension RC. Once Phase 1 is completed, then in Phase 2, items are chosen because their second response variable X_{2ij} maximizes Fisher information along dimension PP. The first response variable is the familiar correct and incorrect response variable X_{1ij} for person *i* ($i = 1, \ldots, I$) and item *j* ($j = 1, \ldots, J$):

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

For the first response variable of Equation 1, a three-parameter logistic (3PL) model was assumed in which $\theta_{RC.i}$ is the person parameter locating person *i* along the RC dimension. Let $\alpha_{RC.j}$, $\beta_{RC.j}$, and *c* be the discrimination, difficulty, and guessing parameters (constrained to be equal across items), respectively, for the RC dimension and item j. We constrained the guessing parameter because when the guessing parameters were allowed to vary, in this and other data, the parameters were tightly clustered around .24, there were some large standard errors, and fit did not improve materially. The model for the RC dimension is the familiar 3PL model:

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

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

$$\begin{aligned} X_{2ij} &= 1 \text{ if person } i \text{ chose the paraphrase incorrect response for item } j,\\ &= 0 \text{ if person } i \text{ chose the elaboration incorrect response for item } j,\\ &= \text{ missing if person } i \text{ chose the correct answer.} \end{aligned} \tag{3}$$

X_{2ij} was modeled using 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 $X_{1ij} = 0$:

$$\pi_{2ij}(X_{2ij} = 1|X_{1ij} = 0) = \frac{\exp[\alpha_{PP.j}(\theta_{PP.i} - \beta_{PP.j})]}{1 + \exp[\alpha_{PP.j}(\theta_{PP.i} - \beta_{PP.j})]}, \tag{4}$$

where $\theta_{PP.i}$ is the location of person i along dimension PP, and $(\alpha_{PP.j}, \beta_{PP.j})$ is the vector of item parameters (discrimination and difficulty) for the PP dimension and item j.

Formulation of the likelihood function requires an assumption of local independence. For this likelihood function, it was assumed that, for any two items j and j', the variables X_{2ij} and $X_{2ij'}$ were 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, $\mathbf{x}_{2i} = [X_{2i1}, X_{2i2}, \ldots, X_{2iJ}]$:

$$L_{2.i} = \prod_{j=1}^{j=J} (\pi_{2ij})^{(1-X_{1ij})X_{2ij}}(1 - \pi_{2ij})^{(1-X_{1ij})(1-X_{2ij})}. \tag{5}$$

Equation 5 has the form of the familiar likelihood function except that each exponent is a product of $(1 - X_{1ij})$. The likelihood function can be maximized by standard software that can properly handle the missing data of Equation 3.

Methods

Item Banks

The item parameters for the simulation study were those of the 360 items in the computerized, but nonadaptive, edition of MOCCA (see Table A.6 in Online Supplementary Material).[1] A 3PL model (Equation 2) was specified for the RC dimension responses. For the RC dimension, the mean and standard deviation of the discrimination ($\alpha_{RC.j}$) parameters were 1.899 and 0.394, and for the difficulty ($\beta_{RC.j}$) parameters, they were −0.217 and 0.440. Guessing parameters (c_j) were fixed at 0.24 for all items based on pilot analyses, in which most items had lower asymptotes near that value.

The 2PL model in Equation 4 was specified for the PP dimension response variable. 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.

The distributions of the β parameters for both dimensions were centered just below zero (see Figures A.1 and A.2 in the Online Supplementary Material), while the majority of the β parameters fell within the range of −1 to 1. This narrow distribution of β parameters led to highly peaked bank information functions. The bank was more informative at the peak for the RC dimension than the PP dimension due to higher α values for the former. However, the lower α values for the PP dimension gave its bank information function a broader shape, leading to more information at the extremes when compared to the RC dimension.

Person θ Parameters

For the person parameters, θ was simulated at 15 discrete values from −2.8 to +2.8 in increments of 0.4, separately for both dimensions; 500 simulated examinees were specified 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, θ values were simulated in this fashion to examine how the dependent variables varied conditional on θ. Because maximum likelihood estimates can be undefined for some response vectors, θ_{RC} (and θ_{PP}) was estimated with weighted maximum likelihood (Guyer & Thompson, 2014; Warm, 1989).

Phase 1: RC Dimension

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. There were two independent variables related to Phase 1: the maximum test length of Phase 1 and the cutoff value of the standard error of measurement (SEM) stopping rule.

Maximum Phase 1 length. The maximum number of items that a student could take in Phase 1 was varied. Three levels of this factor were studied: (1) 40 items, (2) 30 items, and (3) 25 items. In all three conditions, testing in Phase 1 stopped when the estimated SEM reached a specified value or when the number of items reached the specified upper limit.

Stopping rule: SEM. This stopping rule had two levels: 0.35 and 0.30. In the first condition, testing in Phase 1 ended when the SEM reached 0.35 or below or when the number of items reached the maximum Phase 1 length of 25, whichever came first. In the second level of the factor, testing in Phase 1 ended when the SEM reached 0.30 or below or when the number of items reached a maximum Phase 1 length of 25, whichever came first. The estimated SEM was the observed SEM from the simulee's likelihood function.

Dependent variables. Three major dependent variables were examined for Phase 1.

Bias. The first was the average bias in the estimates of θ along dimension RC:

$$\text{Bias}(\theta) = \frac{1}{N}\sum_{i=1}^{i=N}\left(\hat{\theta}_{RC.i} - \theta_{RC.i}\right), \tag{6}$$

that represents the average difference between estimated and generated locations along dimension RC, $\hat{\theta}_{RC.i}$ and $\theta_{RC.i}$, for simulee *i*. Bias conditional on θ_{RC} was also examined.

RMSE. The second dependent variable was the root mean square error:

$$\text{RMSE}(\theta) = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\left(\hat{\theta}_{RC.i} - \theta_{RC.i}\right)^2}. \tag{7}$$

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

Phase 1 Length. Finally, the mean number of items administered to the simulees during Phase 1, that is, average Phase 1 length, was examined.

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 the design, for Phase 1, the between-subjects variable was true (generated) RC dimension ability ($\theta_{RC.i}$), and the within-subjects variable was the manipulated independent variable (e.g., SEM). All ANOVAs were run as a two-way mixed design with $\theta_{RC.i}$ as a between-subjects blocking factor and the independent variable (i.e., maximum Phase 1 length or SEM stopping rule) as a within-subjects factor.

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), was computed 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: PP Dimension

Independent variables. For Phase 2, there were three independent variables: item selection rule, stopping rule, and upper limit on number of items.

Item selection rule: Fisher information versus 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: $I_j\left(\hat{\theta}_{PP.i}\right)$. The second option 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:

$$I_j^*\left(\hat{\theta}_{PP}\right) = [1 - \pi_{1ij}(X_{1ij} = 1)] \times I_j\left(\hat{\theta}_{PP}\right), \tag{8}$$

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

Stopping rule: Confidence interval (CI) versus sequential probability ratio test (SPRT) versus generalized likelihood ratio (GLR). Three stopping rules were compared: a CI rule, the sequential probability ratio test, and the GLR test. As this dimension was conceived, the zero point along the dimension divides persons with a paraphrase PP from those with an elaboration PP.

For the CI 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% CI was computed: CI $=$ $\hat{\theta}_{PP} \pm 1.65\ \hat{s}\left(\hat{\theta}_{PP}\right)$. If the CI included 0, then the algorithm proceeded to select and administer the next item. If the CI did not include 0, the testing stopped. If the CI was below 0, the person was classified as having an elaboration PP. If the CI was above 0, the person was classified as having a paraphrase PP.

The second stopping rule investigated was the SPRT (Thompson et al., 2012; Wang et al., 2021). The SPRT begins by establishing an indifference region along the PP dimension ($-.5$, $.5$) about the cutoff separating paraphrase from elaboration, 0 in the present case. Let UB be the upper bound for the indifference region and let LB be the lower bound: LB < 0 < UB. Let $\mathbf{X}_i = (x_{2i1}, x_{2i2}, \ldots, x_{2ij})$ be the person's response vector after the *j*th item is administered. Two likelihoods are computed: The first is the likelihood of $\mathbf{X}_i$ at $\hat{\theta}_{PP} = \text{UB}$, and the second is the likelihood at $\hat{\theta}_{PP} = \text{LB}$. Let these two likelihoods be designated as $L(UB|\mathbf{X}_i)$ and $L(LB|\mathbf{X}_i)$. After the administration of each item, their ratio is

$$LR = \frac{L(\text{UB}|\mathbf{X}_i)}{L(\text{LB}|\mathbf{X}_i)}. \tag{9}$$

Two cutoffs, *A* and *B,* are then selected, such that *0 < A < B.* If LR < *A*, testing stops, and the person is classified as having an elaboration propensity process. If LR > *B*, then testing stops and the person is classified as having a paraphrase propensity process. If *A* < LR < *B*, then testing proceeds to the next item. For this study, $A = 1/9$ and $B = 9$. In the present application, the SPRT will classify a person as having a paraphrase propensity if the response vector is nine times more likely at the upper bound than at the lower bound. It will classify a person as having an elaboration propensity if the response vector is nine times more likely at the lower bound than at the upper bound.

The third classification rule, the GLR test (Thompson et al., 2012; Wang et al., 2021), employs three likelihoods, $L(\text{UB}|\mathbf{X}_i)$, $L(\text{LB}|\mathbf{X}_i)$, and $L\left(\hat{\theta}_{PP.i}|\mathbf{X}_i\right)$, where $\hat{\theta}_{PP.i}$ is the current maximum likelihood estimate of θ_{PP}. Once an item is administered, a new estimate of $\hat{\theta}_{PP.i}$ is obtained. The GLR equals the ratio in Equation 9 if LB < $\hat{\theta}_{PP.i}$ < UB. If $\hat{\theta}_{PP.i} \geq \text{UB}$, then the algorithm computes

$$\text{GLR} = \frac{L(\hat{\theta}_{PP.i}|\mathbf{X}_i)}{L(\text{LB}|\mathbf{X}_i)}. \tag{10}$$

The numerator of the ratio is the likelihood for $\theta_{PP.i}$ with the maximum likelihood in the interval $\theta_{PP.i} \geq \text{UB}$. If $\hat{\theta}_{PP.i} \leq \text{LB}$, the algorithm computes the ratio

$$\text{GLR} = \frac{L(\text{UB}|\mathbf{X}_i)}{L\left(\hat{\theta}_{PP.i}|\mathbf{X}_i\right)}. \tag{11}$$

The quantity in the denominator of the ratio $L\left(\hat{\theta}_{PP.i}|\mathbf{X}_i\right)$ is the likelihood for the $\theta_{PP.i}$ value with the maximum likelihood in the interval $\theta_{PP,i} \leq \text{LB}$. Testing will stop, and the person will be classified as having an elaboration PP, if GLR < *A*. Testing will stop and the person will be classified as having a paraphrase PP if GLR > *B*. Again, $A = 1/9$ and $B = 9$, the same values used for the SPRT. As we have implemented the GLR (and the SPRT), those for whom the GLR never goes outside the range $1/9 \leq \text{GLR} \leq 9$ are left unclassified. In short, the classification stopping rule factor was a factor with three levels: CI, SPRT, and GLR test.

Stopping rule: Upper limit on number of items. Two different stopping rules for the number of items were studied: 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 person 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. For Phase 2, there were two dependent variables: classification accuracy and Phase 2 length.

Classification accuracy. One of the major dependent variables in Phase 2 was classification accuracy. For measuring classification accuracy, simulees with $\theta_{PP.i} > 0$ were classified as having a true paraphrase propensity, and simulees with $\theta_{PP.i} < 0$ were classified as having a true elaboration propensity. Simulees with $\theta_{PP.i} = 0$ were classified as being members of neither type; therefore, they were not included in calculations of classification accuracy. The proportion of simulees who were correctly classified conditional on true θ_{RC} and θ_{PP} was computed.

Phase 2 length. This dependent variable was operationalized as the mean number of items administered in Phase 2.

Results

Phase 1

Maximum Phase 1 length. The first test termination criterion was the upper bound on the number of items administered (maximum Phase 1 length). The nonadaptive computerized version of MOCCA included 40 items; therefore, with a goal of decreasing the test length of the CAT, upper bounds of 40, 30, and 25 items were compared. There was little effect of the maximum Phase 1 length on bias and RMSE ($\omega^2 < 0.01$ for both), but there was a moderate effect on average test length $\omega^2 = 0.06$ (Figure 3a and b). (Results of the ANOVAs are in Appendix Table A.1 in Online Supplementary Material.) Figure 3c shows that a maximum Phase 1 length of 25 items produced lower average Phase 1 length when compared to maximum Phase 1 lengths of 30 and 40 items for θ_{RC} levels 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.

Stopping rule: Estimated SEM. The main and interaction effects for the stopping rule resulted in $\omega^2 < .001$ (results of the ANOVAs are in Online Supplementary Appendix Table A.2). As shown in Figure 4a and b, a SEM of 0.30 did not result in markedly more accurate results relative to a SEM of 0.35, despite slightly

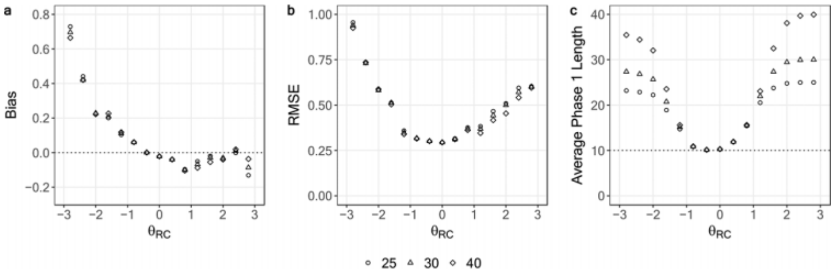


FIGURE 3. *Variation in dependent variables as a function of maximum Phase 1 length (25, 30, and 40 items) conditional on θ_{RC} and standard error of measurement = 0.35.*

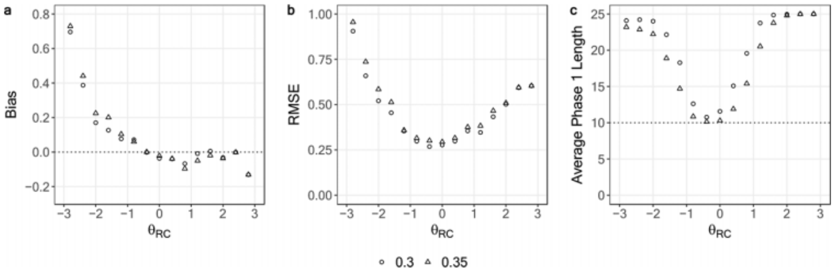


FIGURE 4. *Variation in dependent variables as a function of two standard error of measurement stopping criteria (.30 and .35) conditional on θ_{RC} and maximum Phase 1 length of 25 items.*

increasing the average Phase 1 length at all but the highest true θ_{RC} levels (Figure 4c).

Phase 2

Item selection rule: Fisher information versus weighted Fisher information. When considering item selection rules, it was 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 average Phase 2 length of the CAT. The magnitude of this effect would be dependent upon the underlying θ_{RC} trait value. There were negligible effect sizes for item selection rule and its interaction with θ_{RC} and with θ_{PP} ($\omega^2 < 0.01$; see Online Appendix Table A.3).

Figure 5a shows that weighted Fisher information item selection increased the classification accuracy for $\theta_{RC} > 0$, while also decreasing Phase 2 length for

$\theta_{RC} > -0.4$ (Figure 5b). This effect manifests as relatively uniform increases in accuracy and decreases in Phase 2 length conditional on θ_{PP} (Figure 5c and d). In Figure 5c, there is no data point for classification accuracy at $\theta_{PP} = 0$, because classification accuracy was not defined when $\theta_{PP} = 0$—the true classification at $\theta_{PP} = 0$ was neither paraphrase nor elaboration.

Stopping rule: CI versus sequential probability ratio versus GLR. For classification accuracy, the effect sizes for stopping rule and its interactions were generally small (see Online Appendix Table A.4). Stopping rule had a moderate effect on average Phase 2 length ($\omega^2 = 0.06$), and the interaction between stopping rule and θ_{RC} had a small effect ($\omega^2 = 0.02$); the remaining effects were <0.01.

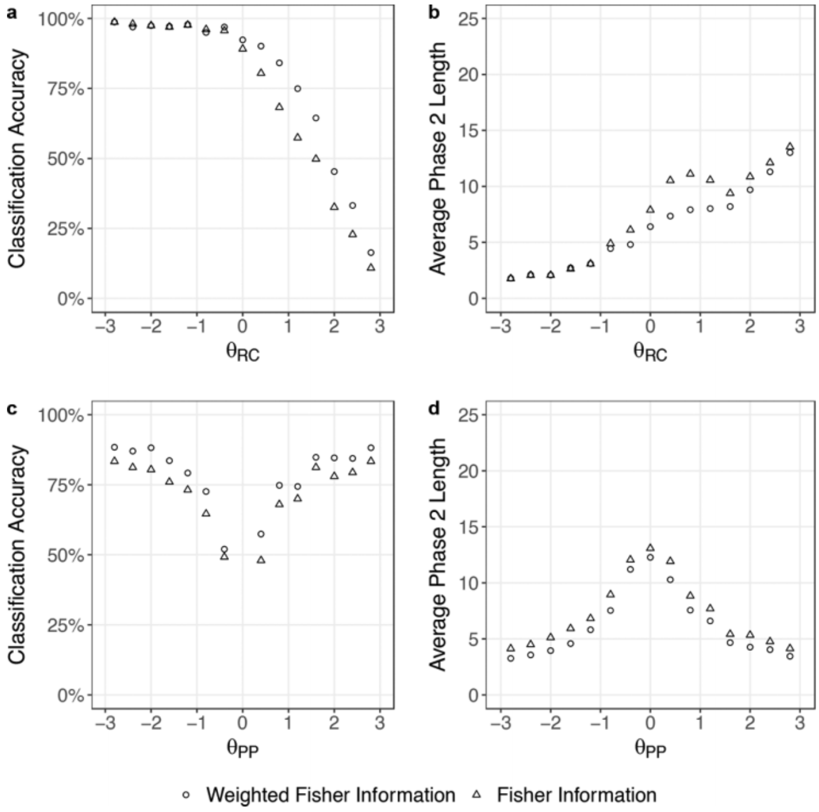


FIGURE 5. *Classification accuracy and average phase 2 length as a function of Fisher information versus weighted Fisher information conditional on* θ_{RC} *(top) and* θ_{PP} *(bottom).*

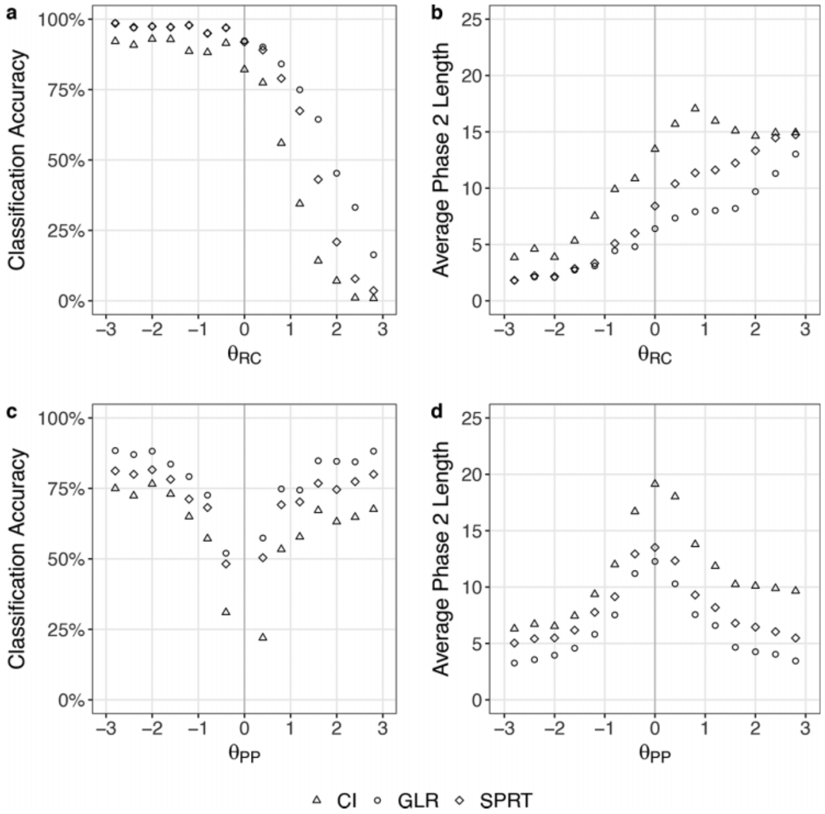


FIGURE 6. *Classification accuracy and average Phase 2 length as a function of classification stopping rule conditional on* θ_{RC} (top) and θ_{PP} (bottom).

Figure 6 presents the comparison of the three classification rules. The GLR and SPRT had near identical accuracy for low θ_{RC}, but GLR had higher accuracy for high θ_{RC} values (Figure 6a). Average Phase 2 length conditional on θ_{RC} (Figure 6b) shows a similar pattern. At all θ_{PP} values, GLR had the best accuracy (Figure 6c) and lowest average Phase 2 length (Figure 6d).

Stopping rule: Maximum Phase 2 length. Two options were considered for the upper limit on Phase 2 length. 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. All maximum length effects resulted in $\omega^2 < 0.01$ (see Online Appendix Table A.5).

Figure 7 shows that because the 25/40 rule allowed for the possible administration of more than 15 items in Phase 2, it led to slightly increased Phase 2

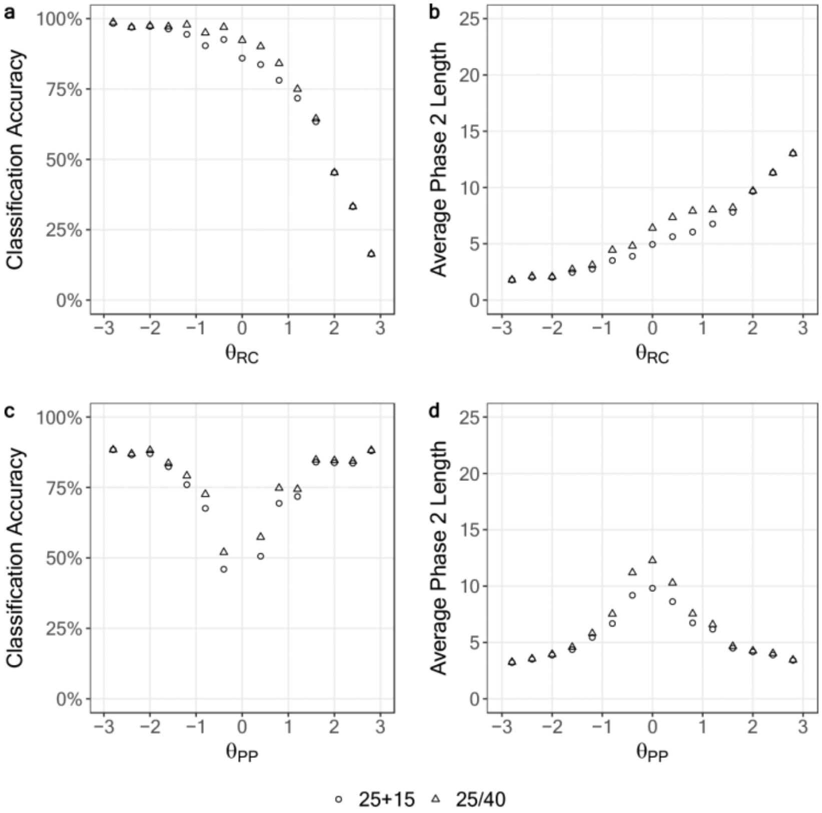


FIGURE 7. *Classification accuracy and average Phase 2 length as a function of maximum Phase 2 length conditional on* θ_{RC} *(top) and* θ_{PP} *(bottom).*

lengths (Figure 7b and d), especially for θ_{RC} and θ_{PP} around zero, and slightly better classification accuracy (Figure 7a and c).

Real Data Example

Following the simulation studies, a pilot version of the MOCCA CAT was administered to 123 third-grade students. There were a few key differences between the pilot administration and the simulation conditions due to (a) developments that occurred between the two studies and (b) various practical considerations for the particular application. First, the pilot administration employed an updated item bank totaling 591 items. The added items were on average more difficult along Dimension RC ($M_{\mathrm{new}} = 0.14$ vs. $M_{\mathrm{orginal}} = -0.22$) and more discriminating along dimension PP ($M_{\mathrm{new}} = 1.51$ vs. $M_{\mathrm{orginal}} = 1.17$) than the original items, both of which were goals during the development of the new items. Second, because MOCCA is primarily intended to assist teachers with providing individualized instruction, for struggling readers in particular, only students with $\hat{\theta}_{RC} < 0$ were administered Phase 2 of the MOCCA CAT and given a classification along dimension PP.[2] The pilot sample turned out to be relatively high-performing, and so only 20 students were administered Phase 2. Third, the starting θ_{RC} value for the CAT was set to $\theta_{RC} = -0.20$ based on the average value for third graders from a previous sample who were administered a nonadaptive MOCCA. Finally, the very first item was randomly selected out of the five most informative items at $\theta_{RC} = -0.20$ to minimize overexposure of the first item.

To compare the precision of the real-data and simulation-data θ_{RC} estimates, Figure 8 shows the observed SEMs from the real data and the average SEMs at each θ_{RC} level in the simulation data. A smoothed line was fit to the real-data SEM data points by locally estimated scatterplot smoothing (LOESS) to facilitate comparison with the average SEMs of the simulated data. The real-data SEMs are consistent with the simulation results in the center of the θ_{RC} continuum. The results appear to differ toward the extremes of the θ_{RC} range, but because there were only three students with $\hat{\theta}_{RC} < -0.7$ and only five with $\hat{\theta}_{RC} > 1.6$, the typical observed SEMs in these ranged are not well estimated.

Figure 9 displays a comparison in the real-data phase lengths and simulation-data average phase lengths, for both Phase 1 (Figure 9a) and Phase 2 (Figure 9b).

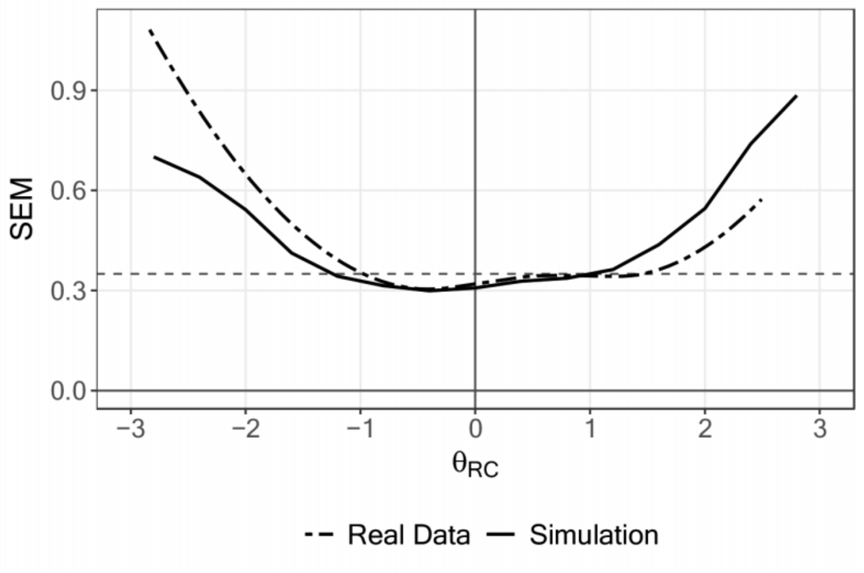


FIGURE 8. *Real-data standard error of measurements (SEMs) and average simulation-data SEMs, conditional on* θ_{RC}.

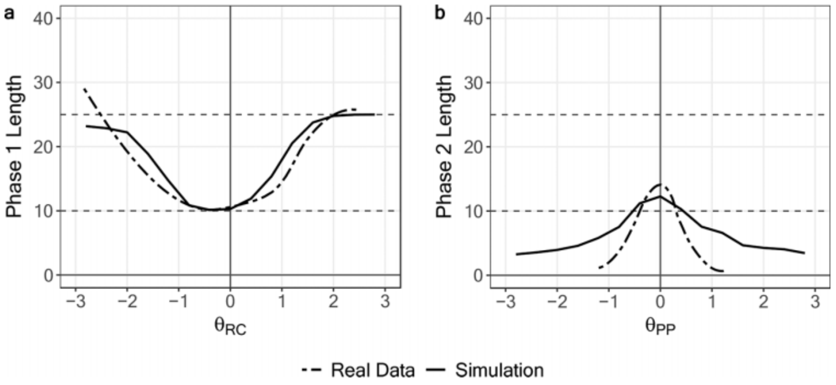


FIGURE 9. *Real-data Phase 1 and 2 lengths and average simulation-data phase lengths, conditional on* θ_{RC} *(for Phase 1) or* θ_{PP} *(for Phase 2).*

The Phase 1 lengths are shown as a function of θ_{RC}, while the Phase 2 lengths are shown as a function of θ_{PP}. Smoothed lines are again fit to the real data by LOESS. The real-data Phase 1 lengths are shorter than the simulation-data average Phase 1 lengths below $\theta_{RC} = -2$ and above $\theta_{RC} = 2$. Because there were only 20 students who were administered Phase 2, the typical Phase 2 lengths are not well estimated. However, it appears that the real-data Phase 2 lengths are shorter than the simulation-data average Phase 2 lengths for θ_{PP} not near zero. Both of these results were expected due to the addition of items with higher θ_{RC} difficulties and larger θ_{PP} discriminations to the item bank used in the real-data pilot study.

It is more difficult to compare the simulation-data and real-data classification results because the primary outcome, classification accuracy, relies on knowledge of the true classification, which is not known in real data. Furthermore, only 20 students were administered Phase 2, which is too small a sample size to make meaningful inferences about the classification performance. Of the 20 students, 11 were classified as paraphrase, nine as elaboration, and zero as indeterminate. The addition of informative items for dimension PP to the bank likely facilitated these high classification rates in the pilot study. Further investigation is needed to determine whether all students who are classified as paraphrase or elaboration have a practically meaningful propensity toward that particular response processes.

Overall, the real-data results are consistent with the simulation results, while also reflecting improvements in the item bank. The pilot version of MOCCA CAT appears to measure students with the required precision on dimension RC, achieve desirable test lengths for the majority of students, and successfully classify struggling readers along dimension PP.

Discussion

The properties of the item bank can have a material effect on the results of CAT simulation studies, so they are frequently based on real item banks. The real items forming the basis for our simulation studies have a notable feature: On both dimensions, item difficulty parameters are concentrated around the center of the dimension, $\theta = 0$, with relatively few items toward the extremes of the distribution, an item bank configuration that is not optimal for the implementation of CATs. This limited variation in item difficulties likely resulted from tight item writing rules constraining item features. The results of this study might not generalize to item banks with more variation in item difficulty. Rather, they are most applicable to item banks composed of items with information functions similar that that used in this study.

A second major feature involves the items on dimension PP. The average item discrimination along dimension PP was lower than that for dimension RC. Furthermore, an item provides information about the PP dimension only when the item is answered incorrectly. For people with very high scores, there are very few incorrect answers from which to infer their location along dimension PP. This led to an interaction whereby classification along dimension PP was more accurate for persons low on dimension RC. This feature, limited information for dimensions beyond the first, is likely a characteristic of tree models. In our application, the classifications are more likely to be used for students low on RC, because they are the ones most in need of supplemental intervention. Consequently, the low end of RC is where classification accuracy is most needed.

Our initial simulation studies did include a simulated item bank with a wider, less peaked distribution of item difficulties and a less peaked information function. This less peaked, simulated item bank improved both measurement and classification accuracy. Because our item writers were able to improve the real item bank but not by as much as we had hoped, we have only reported the results for the narrower item bank more nearly representative of our real item bank.

In Phase 1, two independent variables were examined: maximum number of items (40, 30, and 25) and stopping criterion (.35 vs. 30). Both variables were examined independently rather than in a fully crossed design, with θ_{RC} functioning as a blocking factor. For the maximum-number-of-items independent variable, the effect size for maximum number of items and its interaction with the blocking factor were both small to moderate after controlling for θ_{RC}. The effect of increasing the maximum number of items grew smaller and smaller as the number of items increased. There was little main effect or interaction of maximum test length on bias or RMSE beyond 25 items, controlling for θ_{RC}. Not surprisingly, increasing the maximum number of items had a larger effect on the actual number of items taken, particularly at the extremes of dimension PP, but even that main effect was only of moderate size. These results do not suggest that test length is unimportant but increasing maximum test length beyond

25 increased test length at the extremes with little improvement in bias or RMSE. In a complex, multiphase test, beyond some number of first phase items (25 in the present case), testing time might be better spent on the second phase due to diminishing returns of accuracy with additional items.

In examining the effect of the SEM stopping rule, the total number of items was limited to 25. Results were similar for the stopping rules (.30 vs. .35) as for the maximum number of items. There was little effect of stopping rule or its interaction on bias or RMSE after controlling for θ_{RC}. There was a somewhat larger effect on actual number of items taken, particularly at the extremes of θ_{RC}. Once the SEM reached .35, it improved only very slowly with additional items. At the extremes, it took more items to reach a SEM of .30 with little improvement in bias or RMSE, due (in part) to the nature of the item bank.

Phase 2 focused on two independent variables, item selection rule and classification rule. Classification accuracy and test length were compared after varying the decision rule for item selection, Fisher information and weighted Fisher information. The weighted Fisher information effect was in the predicted direction—higher classification accuracy and shorter test length—but the effects were small after controlling for θ_{RC} or θ_{PP}. Although the effects were small, weighted Fisher information increased classification accuracy while slightly decreasing average test length for those with $\theta_{RC} > 0$.

For the second independent variable, the GLR performed best with the CI performing least well. Effect sizes were not large. GLR both increased classification accuracy and shortened test length. What is notable about the PP dimension results is that there was no trade-off between classification accuracy and test length—the independent variable conditions that maximized classification accuracy (weighted Fisher information and GLR) also decreased test length.

In tree-based sequential multidimensional assessment, CAT offers a way to minimize testing time at the first stage to maximize testing time at later stages. That is, there comes a point where adding additional items improves accuracy along dimension RC only marginally, making additional testing time along dimension RC of limited value. Similarly, with a SEM stopping rule, there will likely be a value of SEM after which accuracy improves too slowly to warrant administering additional items in the first phase. In other applications of similar tree models, simulation research will likely be needed to identify the trade-offs between these test length limits and SEM stopping rules beyond which improvements in accuracy come too slowly. These points will vary depending on the information functions of the item bank.

The results for dimension PP identified a classification rule (GLR) and an item selection function (weighted Fisher information) that both improved classification accuracy and decreased test length, although not by a large amount and not by the same amount at every value of θ_{RC} and θ_{PP}. In tree models, the information function of the testing phases may often be lower for later phases than in Phase 1, either because items have lower discriminations or because not all items administered provide information about dimensions underlying later nodes. Therefore, classification, rather than measurement, may be a more realistic goal. As compared to the alternatives considered here, the GLR classification rule and weighted Fisher information were superior in terms of both classification accuracy and test length.

Psychometrics tends to rely heavily on simulated data research. Seldom is there an attempt to confirm the findings in simulation research with real data. This study provided a limited opportunity to do so. The real data mirrored the major trends of the simulation with some exceptions that can largely be explained by improvements to the item bank that were completed only after the simulation study was completed. The similarities between the trends in the real and simulated data serve to increase confidence in decisions based on the simulated data.

Testing for instruction and diagnosis, rather than rank and sort, requires a major rethinking of testing. It must start with a theory of instruction and diagnosis. Our theory has guided development in several ways. First, because the theory involves identifying struggling readers, the instrument includes a dimension for identifying struggling readers. Second, answer alternatives are designed to facilitate the identification of different types of struggling readers. Third, multiple forms and a large item pool have been developed for progress monitoring, so that a student can take the test multiple times without encountering any item more than once. This, in turn, has led to a statistical model that includes a dimension of overall RC to identify struggling readers and a second dimension to classify struggling readers for the purposes of individualizing additional services. Development of items, response options, and scores have all been guided by a theory of instruction and diagnosis, psychometric theory, statistical simulation results, and real data from several pilot studies. The result is an RC assessment that will, hopefully, prove useful as an outcome measure, as a screening tool, as a progress monitoring instrument, and as an aid for individualizing additional services to struggling readers.

Authors' Note

The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: 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.

ORCID iDs

Mark L. Davison https://orcid.org/0000-0003-3656-9672
Gina Biancarosa https://orcid.org/0000-0002-9471-3145
Patrick Kennedy https://orcid.org/0000-0002-5525-3983

Notes

1. These 360 were the items developed in earlier phases of the project and available as the simulation study began. The current item bank contains nearly 600 items.
2. The effect of this administration procedure was studied via further simulations, not presented here. In summary, they show shorter average test lengths and higher average classification accuracy, since classification accuracy decreases and Phase 2 length increases as θ_{RC} increases.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

Biancarosa, 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. https://doi.org/10.1177/0013164418763255

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, *79*, 403–425.

Cain, K., & Oakhill, J. (2006), Profiles of children with specific reading comprehension difficulties. *British Journal of Educational Psychology*, *76*(4), 683–696. https://doi.org/10.1348/000709905X67610

Carlson, S. E., Seipel, B., & McMaster, K. (2014). Development of a new reading comprehension assessment: Identifying comprehension differences among readers. *Learning and Individual Differences*, *32*, 40–53.

Catts, H. W., Compton, D, Tomblin, J.B., & Bridges, M.S. (2012). Prevalence and nature of late-emerging poor readers. *Journal of Educational Psychology*, *104*(1). https://doi.org/10.1037/a0025323

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, *37*, 419–437. http://doi.org/10.1177/0146621613479818

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

Currie, N. K., & Cain, K (2015). Children's inference generation: The role of vocabulary and working memory *Journal of Child Experimental Psychology, 137*, 57–75.

Davison, M. L., Biancarosa, G., Carlson, S. E., Seipel, B., & Liu, B. (2018). Preliminary findings on the computer administered Multiple-choice Online Causal Comprehension Assessment (MOCCA), a diagnostic reading comprehension test. *Assessment for Effective Intervention*, *43*(3), 169–181. https://doi.org/10.1177/1534508417728685

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

Davison, M. L., Liu, B., Seipel, B., Carlson, S. E., & Biancarosa, G. (2017). *Continuing efforts in selecting an item response theory: Year 2 MOCCA*. Paper presented to the American Educational Research Association.

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

Delmas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28–58.

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

Hermann-Abell, C. F., & DeBoear, G. E. (2011). Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items. *Chemistry Education Research and Practice*, *12*(2), 184–192. https://doi.org/10.1039/C1RP90023D

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141–158. https://doi.org/10.1119/1.2343497

Hong, H., Wang, C., Lim, Y. S., & Douglas, J. (2015). Efficient models for cognitive diagnosis with continuous and mixed-type latent variables. *Applied Psychological Measurement*, *39*(1), 31–43. https://doi.org/10.1177/0146621614524981

Johnson, T. R., & Bolt, D. M. (2010). On the use of factor-analytic multinomial logit item response models to account for individual differences in response style. *Journal of Educational and Behavioral Statistics*, *35*(1), 92–114. https://doi.org/103102/1076998609340529

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272. https://doi.org/10.1177/01466210122032064

Kim, N. (2022). *A mixture IRTree model for understanding response process in rating scales*. Paper presented to the CANAM Online symposium.

Kim, N., & Bolt, D. M. (2021). A mixture IRTree model for extreme responses style: Accounting for response process uncertainty. *Educational and Psychological Measurement*, *81*(1), 131–154. https://doi.org/10.1177/0013164420913915

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. https://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., & Boeck, P. D. (2012). Can fast and slow intelligence be differentiated. *Intelligence*, *40*, 23–32.

Perfetti, C. (2007) Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383, https://doi.org/10.1080/10888430701530730

Pimperton, H., & Nation, K. (2010). Suppressing irrelevant information from working memory: Evidence for domain-specific deficits in poor comprehenders. *Journal of Memory and Language*, *4*, 380–391.

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. https://doi.org/10.1080/10888430701530417

Sadler, P. M. (1998). Psychometric models of student misconceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*, *35*(3), 165–396.

Sireci, S. G. (2022). Six big changes in standardized tests—Including less focus on grading students and more on learning. *Presentation to the National Council on Measurement in Education*. https://theconversation.com/6-big-changes-in-standardized-tests-including-less-focus-on-grading-students-and-more-on-learning-158289

Spencer, M., Wagner, R. K., & Petscher, Y. (2019) The reading comprehension and vocabulary knowledge of children with poor reading comprehension despite adequate decoding: Evidence from a regression-based matching approach. *Journal of Educational Psychology*, *111*(1), 1–14. https://psycnet.apa.org/doi/10.1037/edu0000274

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. https://doi.org/10.1080/08957347.2019.1577247

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287. https://doi.org/10.1037/1082-989X.11.3.287

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

Von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.

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

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. http://doi.org/10.1007/bf02294627

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

Yu, X., Cheng, Y., & Chang, H.-H. (2019). Recent developments in cognitive diagnostic computerized adaptive testing (CD-CAT): A comprehensive review. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models: Models and model extensions, applications, software packages* (pp. 307–331). Springer International Publishing. https://doi.org/10.1007/978-3-030-05584-4_15

Authors

MARK L. DAVISON is a professor in quantitative methods in education, Department of Educational Psychology, University of Minnesota, 56 E. River Road, Minneapolis, MN 55455, USA; e-mail: mld@umn.edu. His research interests include the measurement and prediction of academic achievement.

DAVID J. WEISS is a professor of psychology at the University of Minnesota, 75 E. River Road, Minneapolis, MN 55455, USA; e-mail: djweiss@umn.edu. His research interests include computerized adaptive testing (CAT) and issues in item response theory that support the applications of CAT.

JOSEPH N. DEWEESE is a PhD student in quantitative/psychometric methods, Department of Psychology, University of Minnesota, 75 E. River Road, Minneapolis, MN 55455, USA; e-mail: gree2903@umn.edu. His research interests include item response theory, computerized adaptive testing, and intraindividual change.

OZGE ERSAN was a PhD candidate in quantitative methods in education, Department of Educational Psychology, University of Minnesota, 56 E. River Road, Minneapolis, MN 55455, USA; e-mail: ersan001@umn.edu. She is currently in the Directorate General of Measurement, Evaluation, and Examination Services of the Turkish Ministry of National Education. Her research interests include international large-scale assessments, technology-enhanced items, and computerized adaptive testing.

GINA BIANCAROSA is a professor and the Ann Swindells Chair in Education in the College of Education Department of Special Education and Clinical Sciences at the University of Oregon, 5292 University of Oregon, Eugene, OR 97402, USA; e-mail: ginab@uoregon.edu. Her research interests include the measurement of reading and reading comprehension and understanding growth trajectories in both.

PATRICK C. KENNEDY is a senior research associate at the Center on Teaching and Learning in the College of Education at the University of Oregon, 5292 University of Oregon, Eugene, OR 97403, USA; e-mail: ppaine@uoregon.edu. His research interests center on using quantitative methods, literate programming, and reproducibility to inform the development and improve the instructional utility of educational assessments and interventions.

Manuscript received July 29, 2022
Revision received December 8, 2022
Accepted January 11, 2023