Psychometrika (2025), **90**, 331–357
doi:10.1017/psy.2024.18

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

A Family of Sequential Item Response Models for Multiple-Choice, Multiple-Attempt Test Items

Yikai Lu[1], Jim Fowler[2] and Ying Cheng[1]

[1]Department of Psychology, University of Notre Dame, Notre Dame, IN, USA; [2]Department of Mathematics, The Ohio State University, Columbus, OH, USA

Corresponding author: Ying Cheng; Email: ycheng4@nd.edu

(Received 30 October 2024; accepted 31 October 2024; published online 3 January 2025)

Abstract

We consider a test which allows students to attempt a multiple-choice question multiple times (i.e., multiple attempts). The most extreme form of multiple attempts is the answer-until-correct (AUC) procedure. Previous research has demonstrated that multiple-attempt procedures such as AUC could potentially enhance learning and increase reliability. However, for multiple-choice items, guessing is still non-ignorable. Traditional models of sequential item response theory (SIRT) could model multiple-attempt procedures but fail to take guessing into account. The purpose of this study is to propose SIRT models for multiple-choice, multiple-attempt items (SIRT-MM). First, we defined a family of SIRT-MM models to account for the idiosyncrasies of items, answer options, and examinee behavior. We also explained how these models could improve person parameter estimates by taking into account partial (mis)information of examinees. Second, we conducted model comparisons between the SIRT-MM models, the graded response model, and the nominal response model. Third, we discussed how the item and person parameters can be estimated, and evaluated item and person parameter recovery of SIRT-MM models under different conditions through a simulation study. Finally, we applied the SIRT-MM models to a real dataset and demonstrated their utility through model selection, person parameter recovery, and information functions.

Keywords: answer-until-correct; multiple attempts; multiple choice item; sequential item response theory

1. Introduction

Many tests employ two types of items: constructed-response items and multiple-choice items (Kastner & Stangla, 2011; Lukhele et al., 1994). Constructed response questions require examinees to create their own answers, which can take many forms, including short text responses, an essay, a diagram, an explanation of a procedure, or the step-by-step solution of a mathematical problem (Kastner & Stangla, 2011; Lukhele et al., 1994). Multiple-choice is an item format widely used in testing due to its simplicity of scoring, which consists of answer choices (or alternatives) and in many cases one of them is the correct choice.

Scoring of multiple-choice items can heavily depend on the state of an examinee. The simplest example would be a completely ignorant examinee, who could guess the correct answer choice and possibly receive credit by chance. Specifically, letting K be the total number of answer choices and assuming the examinee does not have any knowledge of a test item, the probability of guessing the correct response, or the expected score when 0/1 (correct-or-incorrect) scoring is used, is $\frac{1}{K}$. We refer to this condition as complete ignorance. An examinee who could eliminate some distractors, or wrong answer choices, will have a higher chance to earn a point. For example, the expected score will be 0.5 when an examinee could eliminate $K-2$ distractors and leave two possible choices including the correct answer choice. Davis (1964) referred to this condition, where an examinee guesses among some, not all, correct and incorrect choices, as partial information (Frary, 1980). On the other hand, Davis (1964) referred to a condition where an examinee is misinformed and eliminates the correct choice, as misinformation (Frary, 1980). The amount of misinformation varies depending on how many choices, including the correct choice, are believed to be incorrect. For instance, if an examinee believes that the correct answer choice and $K-2$ distractors are incorrect, he or she would select the remaining distractor, getting 0 point by 0/1 scoring. On the other hand, if an examinee believes that the actual correct choice is wrong but all the distractors are correct, he or she would select one of the distractors, getting 0 point by 0/1 scoring as well. The former condition is referred to as partial misinformation and the latter condition is referred to as complete misinformation (Frary, 1980). Intuitively, complete misinformation should be penalized more than partial misinformation. Nevertheless, such conditions are handled differently depending on scoring methods. For example, the simple 0/1 scoring method will treat these two misinformation conditions equally, as both examinees would select one of the distractors. However, an answer-until-correct (AUC) procedure, which is a popular multiple-attempt procedure that lets an examinee continue to select answer options until the correct option is picked, tends to give higher expected item scores with lower levels of misinformation (Frary, 1980, 1989; Kane & Moloney, 1978). Frary (1980) gave a good summary of how these two misinformation conditions are handled in various scoring methods including a multiple-attempt procedure, which is the focus of this article.

As a multiple-attempt procedure, AUC has been reported to have various advantages including: (1) AUC can lead to higher reliability than 0/1 scoring by taking into account different levels of examinees' partial (mis)information (Gilman & Ferry, 1972; Hanna, 1975; Slepkov & Godfrey, 2019), (2) AUC could enhance learning by providing immediate corrective feedback between attempts (Epstein et al., 2001), and (3) AUC is strongly preferred by examinees over only one attempt being allowed (DiBattista et al., 2009). Importantly, Epstein et al. (2001) found that their AUC procedure significantly enhances the retention of material from earlier exams. Specifically, in the final exam, students who had previously experienced the AUC approach were twice as likely to answer previously incorrect questions correctly compared to those who had used Scantron forms (Epstein et al., 2001).

Item scoring for a multiple-attempt procedure, including the AUC procedure, can be very simple: $K-u$ where u is the number of attempts an examinee makes. In this way, we can retrieve the levels of partial misinformation by recording the number of attempts. For example, a completely misinformed examinee would continuously select distractors until the last attempt, resulting in zero points, whereas a partially misinformed examinee who believes that one distractor is correct and is unsure about the other choices, would select that distractor at the first attempt and guess from the second attempt on. Thus, the expected score of a partially misinformed examinee would be higher than that of a completely misinformed examinee. A multiple-attempt procedure can also take into account the different levels of partial information. For example, if a partially informed examinee could eliminate a number of distractors and leave s remaining choices, they are guaranteed to have a score of $K-s$ or better. Moreover, different item scoring schemes are possible. Slepkov and Godfrey (2019) conducted analyses of the reliability of several multiple-attempt tests using different item scoring schemes. In Slepkov and Godfrey (2019), the most popular scoring scheme is one that grants full credit if the first attempt was successful, half credit if the second attempt is successful, one-tenth credit if the third attempt was successful, and no credit otherwise.

These scoring schemes are based on classical test theory. Classical test theory is a class of measurement models that are based on the total sum of item scores, and typically each item is scored by the 0/1 scoring. When such scoring schemes for a multiple-attempt procedure are used, we calculate the total sum of item scores as an estimate of the ability of an examinee. Another approach to model

https://doi.org/10.1017/psy.2024.18 Published online by Cambridge University Press

the ability of an examinee is to use item response theory (IRT; De Ayala, 2009). Tutz (1990) proposed sequential item response (SIRT) models, which are motivated by "a genuine stepwise approach" to emphasize its difference from the partial credit model. Unlike the latter, SIRT models can model a person's consecutive attempts at an item, such as a test of psychomotor skills. One advantage of SIRT models is that item parameters specific for future attempts do not affect ability estimation at the earlier attempts (Tutz, 1990). Thus, SIRT models could be used for modeling a multiple-attempt procedure that allows an unlimited or limited number of attempts.

More specifically, following the notations in Culpepper (2014), SIRT models could be formulated as follows. Suppose a test item has a multiple-attempt procedure that allows an examinee to submit answers until they reach the correct answer. Let X_j be a random variable representing the number of attempts an examinee needed to submit a correct answer on item j and Y_{ju} represents a Bernoulli random variable of whether the examinee submitted a correct or incorrect response on attempt u. In SIRT, $P(X_j = u|\theta)$ where $u = 1,2,\dots$ could be constructed by assuming $P(Y_{ju} = 1|Y_{j(u-1)} = 0) = H_j(\theta,u)$ and

$$\begin{aligned}P(X_j = u|\theta) &= P(Y_{ju} = 1|Y_{j(u-1)} = 0,\theta)P(Y_{j(u-1)} = 0|Y_{j(u-2)} = 0,\theta)\\ &\quad \dots P(Y_{j2} = 0|Y_{j1} = 0,\theta)P(Y_{j1} = 0|\theta)\\ &= H_j(\theta,u)\prod_{k=1}^{u-1}[1-H_j(\theta,k)].\end{aligned} \tag{1}$$

We often assume that $H_j(\theta,u)$ is a function of item parameters, $H_j(\theta,u) = H(\theta,\Omega_{ju})$, where Ω_{ju} is item parameters for item j at attempt u. By assuming $H(\theta,\Omega_{ju})$ to be a Rasch model,

$$H_j(\theta,u) = H(\theta,\Omega_{ju}) = \frac{\exp(\theta - b_{ju})}{1+\exp(\theta - b_{ju})}, \tag{2}$$

where $b_{ju} \in \Omega_{ju}$, we get the Rasch sequential item response model, which was proposed by Tutz (1990).

However, one problem in using the Rasch sequential item response model for multiple-choice, multiple-attempt test items is that it does not take into account guessing at each attempt, which can be non-trivial especially at later attempts when some answer options have already been eliminated. While existing SIRT models are suitable for a multiple-attempt procedure with constructed responses, an appropriate model is yet to be proposed for a multiple-attempt procedure with multiple-choice items. Thus, the goal of this study is to formally propose new SIRT models, which we call the "SIRT-MM" models (MM stands for multiple-choice, multiple-mttempt), to effectively score multiple-attempt responses for multiple-choice test items. This will be achieved by taking into account the structure of a multiple-choice test item, especially considering the homogeneity or heterogeneity of distractors and the process of elimination of answer choices after reattempts. As a result, we will address the issue of guessing at each attempt. Subsequently, we will also (1) evaluate parameter recovery under various test length and sample size conditions, (2) compare SIRT-MM models with competing models such as the graded response model and the nominal response model for multiple attempts data, and (3) demonstrate the usage of SIRT-MM models using real data.

2. Methods

2.1. *The basic SIRT-MM model*

In this section, we will suppress the subscript j denoting individual items for simplicity (e.g., denoting $H_j(\theta,u)$ as $H(\theta,u)$). Theoretically, to model items using SIRT models, we can design any function for $H(\theta,u)$, and thus an infinite number of the variants of SIRT models could be created. In our context, we need to consider a $H(\theta,u)$ suitable for multiple-choice test items. We begin by considering the structure of a multiple-choice item. Suppose a multiple-choice item has K choices, including one correct choice and $K-1$ distractors, and its choice set as $S = \{v_1, v_2, \dots, v_K\}$, which is a set of all K answer choices of a multiple-choice item. We allow K attempts because any examinee could reach a correct choice by the Kth attempt by eliminating all the distractors. Technically speaking, we only need $K-1$ attempts. However, for the sake of the clarity of our models, we will include the Kth attempt as a response category to differentiate whether the $(K-1)$th attempt is successful or not.

We model multiple-choice, multiple-attempt test items following the discrete choice theory (Agresti, 2013; Ben-Akiva, 1985; Benson et al., 2016; Luce, 1959). The fundamental principle of discrete choice analysis is utility maximization, which assumes that a decision maker selects the option or alternative with the highest utility among all available alternatives at the time (Ben-Akiva, 1985). In the testing context, each answer option of a multiple-choice test item is considered an alternative, and an examinee's perceived correctness of an option is its utility. Utility maximization means an examinee would always try to choose an answer option with the highest perceived correctness.

However, the deterministic view of the choice theory has a limitation in modeling examinees' behavior using a single latent variable θ, because θ cannot fully explain the variations of item responses and there always is some randomness not captured. Therefore, we will adopt the probabilistic choice theory for modeling examinees' behavior.

In probabilistic choice theory, the "choice axiom" (Ben-Akiva, 1985; Luce, 1959) states that the probability of choosing any answer choice v from the choice set S would satisfy

$$P(v|S) = P(v|\tilde{S} \subset S) = P(v|\tilde{S})P(\tilde{S}|S), \tag{3}$$

where $\tilde{S}$ is any subset of S and $P(\tilde{S}|S)$ is the probability of choosing an answer choice in $\tilde{S}$. The choice axiom suggests that if some distractors are removed from the choice set, the relative probabilities for the remaining options are unchanged (Ben-Akiva, 1985). The "choice axiom" implies *independence from irrelevant alternatives* (IIA; Luce, 1959):

$$\frac{P(v_a|S)}{P(v_b|S)} = \frac{P(v_a|\tilde{S})}{P(v_b|\tilde{S})}, \tag{4}$$

which suggests that the odds of choosing v_a over v_b do not depend on the other options in the choice set (Agresti, 2013).

The IIA assumption is widely used and discussed in statistics literature (Agresti, 2013; Benson et al., 2016), such as in multinomial logit models, e.g., multinomial logistic regression (Agresti, 2013; Ben-Akiva, 1985), and other "divide-by-total" models (Thissen & Steinberg, 1986), e.g., (generalized) partial credit model (Masters, 1982; Muraki, 1992) and nominal response model (Bock, 1972). In fact, the divide-by-total models can be derived under the IIA assumption. In other words, when IIA holds, a utility model of $P(v|\tilde{S})$ for any $\tilde{S} \subseteq S$ will be

$$P(v|\tilde{S}) = \frac{w(v)}{\sum_{v' \in \tilde{S}} w(v')}, \tag{5}$$

where a utility measure for answer choice v is represented as a positive real valued function $w(v)$, which is directly proportional to the choice probability. In modeling a multiple-choice test item, utility measure can be considered a function of the latent ability of an examinee, θ. Here, we take as the utility measure the probability of the option v being perceived as true by an examinee with ability level of θ, i.e., $w(v) = p_v(\theta)$. As we assume the IIA, $w(v)$ does not change after eliminating any other answer choice.

In sum, the IIA assumption implies that (1) the probability of making a choice can be expressed as a utility (or divide-by-total) model and thus gets re-scaled proportionally at every attempt. In other words, the probability of choosing an answer option is simply a scaling constant away from the utility measure of the option, and can thus be treated as interchangeable; and (2) eliminating an answer choice does not change the utilities of other alternatives, resulting in the unchanged relative choice probabilities. In the context of multiple attempts, we call the latter implication *attempt invariance*, which means that the utility measures will not change over attempts. This is a reasonable assumption as long as after every attempt, feedback is only given regarding their previous choice being correct or incorrect, without any additional information about the remaining answer choices. Later when we relax the attempt invariance

https://doi.org/10.1017/psy.2024.18 Published online by Cambridge University Press

assumption, we express the probability as $p_v(\theta, u)$ at attempt u, indicating that the probability depends on the number of attempts having been made.

To formulate the SIRT-MM model, we assume that $p_v(\theta)$ could be sufficiently modeled by a single latent variable θ and item parameters. Let T be the correct answer choice and $p_T(\theta)$ be the probability of considering the correct answer choice to be true. Let D_u be the distractor with the u-th largest utility for each examinee and $p_{D_u}(\theta)$ be the probability of endorsing the distractor at the u-th attempt, where each examinee would select in the order of $D_1, D_2, \ldots, D_{K-1}$ when they consistently make failed attempts. Note that we are not interested in specific choices of distractors for the ordering of $D_1, D_2, \ldots, D_{K-1}$ for each examinee, but we are only interested in modeling the expected u-th highest utility of a distractor at given θ (i.e., $p_{D_u}(\theta)$), assuming that each examinee selects the answer choices with the highest utility subjectively judged by them.

Recall that, in SIRT, $P(X = u|\theta)$ where $u = 1, 2, \ldots, K$ could be constructed by assuming $P(Y_u = 1|Y_{u-1} = 0, \theta) = H(\theta, u)$ and

$$\begin{aligned} P(X = u|\theta) &= P(Y_u = 1|Y_{u-1} = 0, \theta)P(Y_{u-1} = 0|Y_{u-2} = 0, \theta) \\ &\quad \ldots P(Y_2 = 0|Y_1 = 0, \theta)P(Y_1 = 0|\theta) \\ &= H(\theta, u)\prod_{k=1}^{u-1}[1 - H(\theta, k)]. \end{aligned} \tag{6}$$

Then, based on the utility model presented in Eq. (5), the probability of submitting a correct answer on the first attempt is

$$H(\theta, 1) = P(Y_1 = 1|\theta) = \frac{p_T(\theta)}{\sum_{v=1}^{K} p_v(\theta)} = \frac{p_T(\theta)}{p_T(\theta) + \sum_{k=1}^{K-1} p_{D_k}(\theta)}. \tag{7}$$

The conditional probability of submitting a correct answer at the second attempt is

$$H(\theta, 2) = P(Y_2 = 1|Y_1 = 0, \theta) = \frac{p_T(\theta)}{[\sum_{v=1}^{K} p_v(\theta)] - p_{D_1}(\theta)} = \frac{p_T(\theta)}{p_T(\theta) + \sum_{k=2}^{K-1} p_{D_k}(\theta)}, \tag{8}$$

where a distractor D_1 is initially mistakenly selected. This supports an intuition that $H(\theta, u)$ will be higher as u gets larger by eliminating distractors.

We begin by deriving the simplest form of an SIRT-MM model. This simple model assumes that all the distractors will are equally appealing to examinees, even after reattempts. In other words, we assume that all the distractors have the same probability of being selected given θ (i.e., homogeneity of distractors). Mathematically put, $p_{D_1}(\theta) = p_{D_2}(\theta) = \ldots = p_{D_{K-1}}(\theta)$ and we can simply denote $p_{D_u}(\theta)$ as $p_D(\theta)$. One advantage of assuming both IIA and homogeneity of distractors is that it allows us not to assume a shape for $H(\theta, u)$ for $u = 2, \ldots$. In fact, $H(\theta, u)$ for $u = 2, \ldots$ could be analytically derived from $H(\theta, 1)$. Therefore, by assuming the shape of $H(\theta, 1)$ to be a 3PL logistic function, as the first attempt is technically the same as the 0/1 scoring, the whole model could be derived. Later in this article, we introduce parameterizations that allow us to relax both assumptions.

Now, for $u = 1, 2, \ldots, K$, $H(\theta, u)$ can be instead written as

$$H(\theta, u) = \frac{p_T(\theta)}{p_T(\theta) + \sum_{k=u}^{K-1} p_{D_k}(\theta)} = \frac{p_T(\theta)}{p_T(\theta) + (K - u)p_D(\theta)}. \tag{9}$$

We can observe that the reciprocal of this probability, $\frac{1}{H(\theta, u)}$, decreases linearly by $\frac{p_D(\theta)}{p_T(\theta)}$ as u increases since:

$$\frac{1}{H(\theta, u)} = 1 + (K - u)\frac{p_D(\theta)}{p_T(\theta)}. \tag{10}$$

Finally, by assuming $H(\theta, 1)$ to be a 3PL logistic function with a fixed pseudo-guessing parameter of $\frac{1}{K}$, which we denote as the "2.5 PL" model as it is between the 2PL and 3PL models

(Bizot & Goldman, 1994),

$$H(\theta,1) = \frac{1}{K} + \left(1 - \frac{1}{K}\right)\frac{\exp(a(\theta-b))}{1+\exp(a(\theta-b))} = \frac{p_T(\theta)}{p_T(\theta)+(K-1)p_D(\theta)}. \tag{11}$$

Thus,

$$\begin{aligned}\frac{p_D(\theta)}{p_T(\theta)} &= \frac{\frac{1}{H(\theta,1)} - 1}{K-1} \\ &= \left[\frac{K+K\exp(a(\theta-b))}{1+K\exp(a(\theta-b))} - 1\right]/[K-1] \\ &= \frac{1}{1+K\exp(a(\theta-b))}\end{aligned} \tag{12}$$

and

$$\begin{aligned}\frac{1}{H(\theta,u)} &= 1 + (K-u)\frac{p_D(\theta)}{p_T(\theta)} \\ &= 1 + \frac{K-u}{1+K\exp(a(\theta-b))}.\end{aligned} \tag{13}$$

It is worth noting that we only model $\frac{p_D(\theta)}{p_T(\theta)}$ in the equations, instead of directly modeling $p_D(\theta)$ and $p_T(\theta)$, respectively. Finally, the unconditional probability of choosing the correct choice for item *j* at the u-th attempt is derived as follows:

$$\begin{aligned}P(X=u|\theta) &= H(\theta,u)\prod_{k=1}^{u-1}[1-H(\theta,k)] \\ &= \frac{(K-1)![1+K\exp(a(\theta-b))]}{(K-u)!\prod_{k=1}^{u}[K-k+1+K\exp(a(\theta-b))]},\end{aligned} \tag{14}$$

which is the simplest SIRT-MM model.

Figure 1 shows example item category response functions (i.e., $P(X=u|\theta)$) when $a=1.7, b=0.0$, $K=5$. As expected, we can observe that at any θ, $\sum_{u=1}^{K} P(X=u|\theta) = 1$, and $P(X=1|\theta)$ (or $u=1$) has the same shape as a 3PL logistic model. This figure also shows that conditioning on $X=2$ (or $u=2$), the middle range of θ is the most likely when $b=0$. This is intuitive as those who need exactly two attempts to get the right answer likely do not have very low or high θ. Since we assume a fixed pseudo-guessing parameter of $\frac{1}{K}$, $P(X=u|\theta)$ converges to $\frac{1}{K}$ as $\theta \to -\infty$, suggesting complete ignorance will occur as $\theta \to -\infty$. This figure also shows that $P(X=1|\theta)$ is the highest among all $P(X=u|\theta)$ at any θ. To allow $P(X=u|\theta)$ for $u=2$ or above to be larger than $P(X=1|\theta)$ for some θ, we need to relax the homogeneity of distractors assumption and attempt invariance.

2.2. More general SIRT-MM models

Now, we turn to a more general case where distractors are not homogeneous, in particular, one distractor being the most attractive. Consider examinees with ability θ who try to evaluate all answer options of two test items. Let d_k be the distractor k of an item, specified by its position within the item. Note that d_k is different from D_u, which we introduced earlier to represent the distractor with the u-th largest utility for each examinee. Suppose on average examinees with ability θ perceive the chance for the four options of item 1 being the correct choice as $(p_T(\theta), p_{d_1}(\theta), p_{d_2}(\theta), p_{d_3}(\theta)) = (0.25, 0.25, 0.25, 0.25)$, and that of item 2 as $(p_T(\theta), p_{d_1}(\theta), p_{d_2}(\theta), p_{d_3}(\theta)) = (0.25, 0.65, 0.05, 0.05)$. Table 1 presents the probabilities of submitting a correct response at each attempt for the two items based on the utility model. Obviously, the probability of submitting a correct response at the first attempt is 0.25 for both

https://doi.org/10.1017/psy.2024.18 Published online by Cambridge University Press

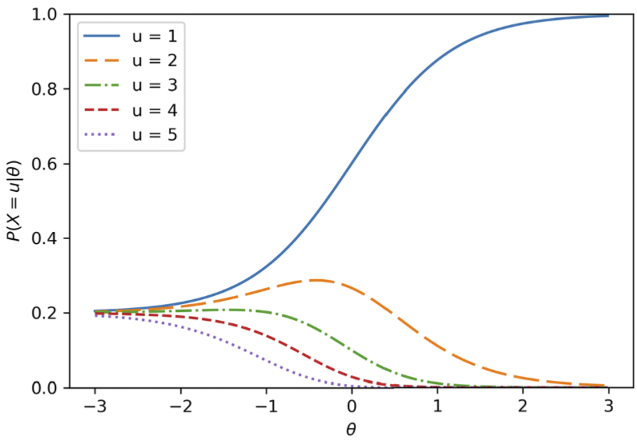


Figure 1. Item category response function: $a = 1.7, b = 0.0, K = 5$.

Table 1. Probabilities of submitting a correct response at each attempt for two hypothetical test items

Item with $(p_T(\theta), p_{d_1}(\theta), p_{d_2}(\theta), p_{d_3}(\theta))$	Condition	Attempt			
		1	2	3	4
$(0.25, 0.25, 0.25, 0.25)$	Complete ignorance	0.25	0.25	0.25	0.25
$(0.25, 0.65, 0.05, 0.05)$	Partial (mis)information	0.25	0.49	0.21	0.05

items. However, the chance of submitting a correct response at the second attempt is different. For the the first item, it is $(1-0.25)\cdot\frac{0.25}{0.25+0.25+0.25} \approx 0.25$. For the second item it is $0.65 * \frac{0.25}{0.25+0.05+0.05} + 2 * 0.05 * \frac{0.25}{0.25+0.65+0.05} \approx 0.49$. Note that the first term of the second equation, which calculates the probability of choosing d_1 first and then the correct answer choice, is $0.65 * \frac{0.25}{0.25+0.05+0.05} \approx 0.46$, meaning that when an examinee makes two attempts for the second item, they are likely tricked by the most attractive distractor, d_1, and select d_1 at the first attempt. These results also suggest that a multiple-attempt procedure penalizes complete ignorance more than partial (mis)information at the second attempt. For the first item, the person considers the correct answer choice to be equally likely, while for the second item, the person at least believes that the correct answer choice is more probable than d_2 and d_3. In the second case, partial information helps avoid needing more than two attempts to answer the item correctly.

We also consider another general case where the utility of any answer choice will change over reattempts. Suppose that the utility of answer choice v at attempt u is $p_v(\theta, u)$. The IIA assumption implies attempt invariance, which means that $p_v(\theta,1) = p_v(\theta,2) = \ldots = p_v(\theta,K)$, allowing us to denote $p_v(\theta,u)$ as $p_v(\theta)$. However, this could be a strong assumption in a multiple-attempt procedure because the population changes after reattempts and specific characteristics of items might affect changes in $p_v(\theta,u)$ over reattempts. For example, a test item that requires certain factual knowledge (e.g., trivia questions) to answer might divide examinees into those who know the answer with confidence and those who do not know the answer at all. In such a case, conditioning on some θ, the population could be divided into two groups by whether they know the requisite fact. For some value of θ, the first group of examinees might believe $(p_T(\theta), p_{d_1}(\theta), p_{d_2}(\theta), p_{d_3}(\theta)) = (0.91, 0.03, 0.03, 0.03)$, while the second group of examinees might believe $(p_T(\theta), p_{d_1}(\theta), p_{d_2}(\theta), p_{d_3}(\theta)) = (0.25, 0.25, 0.25, 0.25)$. If there is an equal number of examinees from each group in the population, on average, $p_T(\theta)$ would be around 0.58 at the first attempt. However, after the first attempt, most likely only the latter group of

https://doi.org/10.1017/psy.2024.18 Published online by Cambridge University Press

examinees would proceed to the second attempt, leading to a much lower average $p_T(\theta)$ at the second attempt. A similar phenomenon was documented by Lyu et al. (2023)) in which they explain that certain item characteristics can have a larger effect after reattempts, resulting in higher difficulty estimates for multiple-attempt items.

To accommodate scenarios described above, we can formulate a more general SIRT-MM model which relaxes both the homogeneity of distractors and the attempt-invariance assumptions by introducing more parameters to vary the average utility of all the unattempted distractors relative to that of the correct answer choice across different attempts.

Recall that $H(\theta,u)$ can be expressed as follows

$$H(\theta,u) = \frac{p_T(\theta,u)}{p_T(\theta,u) + \sum_{k=u}^{K-1} p_{D_k}(\theta,u)}. \tag{15}$$

Thus,

$$\frac{1}{H(\theta,u)} = 1 + \sum_{k=u}^{K-1} \frac{p_{D_k}(\theta,u)}{p_T(\theta,u)}. \tag{16}$$

To model $H(\theta,u)$, we model the average of all the unattempted distractors at the uth attempt[1]:

$$\overline{p_D}(\theta,u) = \frac{1}{K-u} \sum_{k=u}^{K-1} p_{D_k}(\theta,u). \tag{17}$$

Then,

$$\begin{aligned}\frac{1}{H(\theta,u)} &= 1 + \sum_{k=u}^{K-1} \frac{p_{D_k}(\theta,u)}{p_T(\theta,u)} \\ &= 1 + (K-u)\frac{\overline{p_D}(\theta,u)}{p_T(\theta,u)}.\end{aligned} \tag{18}$$

In modeling $\frac{\overline{p_D}(\theta,u)}{p_T(\theta,u)}$, as an extension of the simplest SIRT-MM model, which assumes

$$\frac{p_D(\theta)}{p_T(\theta)} = \frac{1}{1 + K\exp(a(\theta - b))}, \tag{19}$$

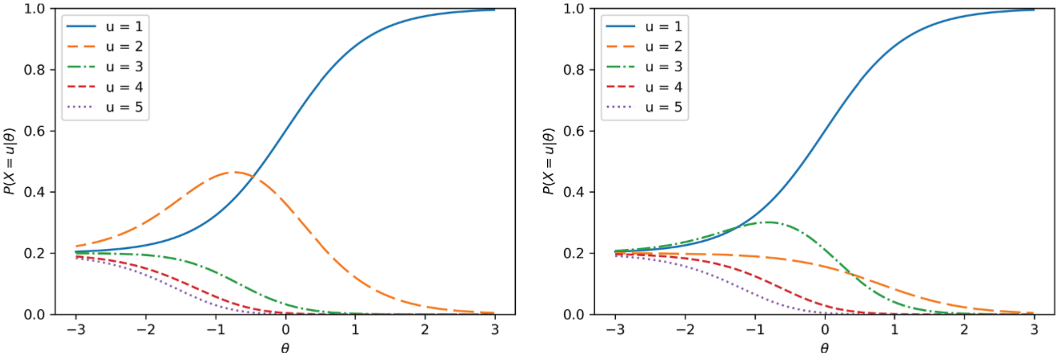
we propose to introduce the attempt-specific "difficulty-shift" parameter $\gamma_u \in \mathbb{R}$ for $u = 2, \ldots, K$ for a more general SIRT-MM model:

$$\frac{\overline{p_D}(\theta,u)}{p_T(\theta,u)} = \frac{1}{1 + K\exp(a(\theta - b + \gamma_u))}. \tag{20}$$

Therefore,

$$\begin{aligned}\frac{1}{H(\theta,u)} &= 1 + (K-u)\frac{\overline{p_D}(\theta,u)}{p_T(\theta,u)} \\ &= 1 + \frac{K-u}{1 + K\exp(a(\theta - b + \gamma_u))}.\end{aligned} \tag{21}$$

[1] One approach to model $H(\theta,u)$ is to model $\frac{p_{D_k}(\theta)}{p_T(\theta)}$ directly as a function of item parameters and θ. However, an issue of this parameterization is that $H(\theta,u)$ will depend on attempt-specific parameters from later attempts, which makes it not an SIRT model anymore. Also, we cannot estimate the model when we do not have a large sample size or we set the maximum number of attempts to be less than K where later attempts are not available.



(a) $\gamma_2 = 1$ (b) $\gamma_2 = -1$

Figure 2. Item category response function: $a = 1.7, b = 0.0, \gamma_3 = 0.5, \gamma_4 = 0$, and $K = 5$ with different γ_2.

This leads to:

$$\begin{aligned} P(X = u|\theta) &= H(\theta,u) \prod_{k=1}^{u-1} [1 - H(\theta,k)] \\ &= \frac{(K-1)![1 + K\exp(a(\theta - b + \gamma_u))]}{(K-u)!\prod_{k=1}^{u}[K-k+1+K\exp(a(\theta - b + \gamma_k))]}, \end{aligned} \tag{22}$$

where $\gamma_1 \equiv 0$ and $\gamma_K \equiv 0$. This is the more general SIRT-MM model, of which the simplest SIRT-MM model (Eq. (14)) is a special case.

We define non-zero γ_u parameters only for $u = 2, \ldots, K-1$ because: (a) γ_1 will lead to over-parameterization due to the existence of b, and (b) γ_K is not necessary since only one answer choice will be left after the $K-1$th attempt. Note that γ_u is item and attempt specific, but does not vary across examinees. We could further relax $\frac{\overline{p_D}(\theta,u)}{p_T(\theta,u)}$ by allowing a to vary over each attempt at u (i.e., modeling a_{ju} or $(a + \delta_{ju})$). We discuss this extension in the Supplementary Material and discussion section.

2.3. *Interpreting γ parameters*

The simplest interpretation of γ_u is that γ_u regulates the probability of making a successful attempt at the uth attempt. Specifically, when γ_u increases, $P(X = u|\theta)$ increases. Similarly, when γ_u decreases, $P(X = u|\theta)$ is decreased.

Figure 2 shows example item category response functions (ICRFs) when $a = 1.7, b = 0.0, K = 5, \gamma_3 = 0.5, \gamma_4 = 0$ and two different γ_2s. The left panel shows the ICRF with $\gamma_2 = 1$ and the right panel shows the ICRF with $\gamma_2 = -1$. All parameters except for the γ parameters are the same as those for Figure 1. Note that $P(X = 1|\theta)$ is unaffected by any γ_u parameters, as non-zero γ_u are only defined for $u = 2, \ldots, K-1$, which could not influence $u = 1$. The major difference between the two panels lies in $P(X = 2|\theta)$, which has a pronounced peak around $\theta = -0.5$ when $\gamma_2 = 1$ and is rather flat around $\theta = -0.5$ when $\gamma_2 = -1$. As a result, $P(X = u|\theta)$ for $u > 2$ are also affected accordingly, which are smaller when $\gamma_2 = 1$ and larger when $\gamma_2 = -1$. This indicates that examinees with lower ability (e.g., $\theta < -1$) are more likely to require only two attempts to answer correctly when $\gamma_2 = 1$, whereas they are more likely to require three attempts when $\gamma_2 = -1$. Therefore, by adjusting γ parameters, different types of item category response functions can be captured.

More specifically, γ_u governs the change of probability ratio between the average of the unattempted distractors and the correct answer option and at the uth attempt (i.e., $\frac{\overline{p_D}(\theta,u)}{p_T(\theta,u)}$) compared to the first attempt.

https://doi.org/10.1017/psy.2024.18 Published online by Cambridge University Press

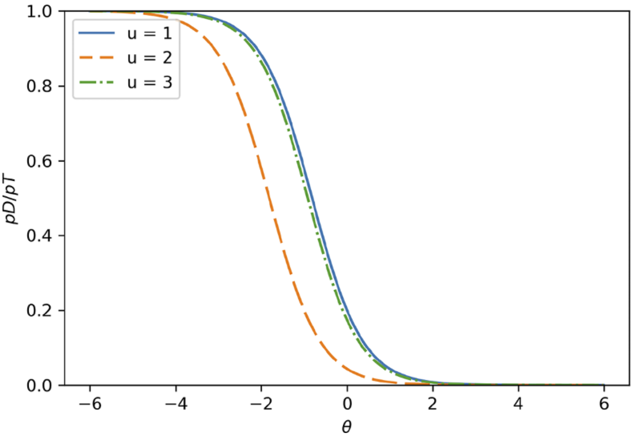


Figure 3. $\frac{\overline{p_D}(\theta,u)}{p_T(\theta,u)}$ when $a = 1.7, b = 0.0, \gamma_2 = 1, \gamma_3 = 0.1, K = 4$.

Figure 3 shows $\frac{\overline{p_D}(\theta,u)}{p_T(\theta,u)}$ when $a = 1.7, b = 0.0, K = 4, \gamma_2 = 1, \gamma_3 = 0.1$. We set $\gamma_3 = 0.1$ instead of $\gamma_3 = 0$ to prevent $u = 1$ and $u = 3$ lines from overlapping. We can observe that $\frac{\overline{p_D}(\theta,2)}{p_T(\theta,2)}$ is shifted to the left by $\gamma_2 = 1$ compared to $\frac{\overline{p_D}(\theta,1)}{p_T(\theta,1)}$. Similarly, $\frac{\overline{p_D}(\theta,3)}{p_T(\theta,3)}$ is shifted to the left by $\gamma_2 = 0.1$ compared to $\frac{\overline{p_D}(\theta,1)}{p_T(\theta,1)}$

2.4. *Possible factors that affect γ parameters*

Increasing γ_u parameters over attempts can be caused by the heterogeneity of distractors. This means when $\gamma_2 > 0$, $p_{D_1} > p_{D_2}$ is expected. Furthermore, when $\gamma_3 > \gamma_2$, $p_{D_2} > p_{D_3}$ is expected. For example, when $\gamma_2 = 1.0, \gamma_3 = 1.0$, and $K = 4$ for an item, there is at least one attractive distractor that will make examinees more likely to make two attempts. Numerically, a positive γ_u reduces $\frac{\overline{p_D}(\theta,u)}{p_T(\theta,u)}$ at the uth attempt. Logically, increasing γ_u indicates knowledge gain through correcting partial misinformation because after a failed attempt, a distractor with high utility will be eliminated, leading to $\frac{\overline{p_D}(\theta,u)}{p_T(\theta,u)}$ to be smaller at later attempts and the correct answer option even more appealing. Therefore, increasing γ_u could signify the heterogeneity of distractors. Increasing γ_u might also be caused by informative feedback such as hints after a wrong response.

On the other hand, as we described earlier, decreasing γ_u parameters can be caused by the population changes after reattempts and specific characteristics of items. In the example of an item requiring factual knowledge, p_T could decrease because the examinees who fails the first attempt would likely fail the second attempt as well, and they represent the majority of those who need the second attempt. Thus, $\frac{\overline{p_D}(\theta,u)}{p_T(\theta,u)}$ would increase from the first to second attempt, which would be captured by a negative γ_2.

In addition, negative γ parameters could result from having a large number of examinees who are being inattentive and fail to eliminate already selected distractors in reattempts. In this article, we assume that examinees are attentive. However, if the system allows examinees to select the same wrong answer option repeatedly, negative γ_u could result.

2.5. *Setting the maximum number of attempts*

One advantage of using an SIRT model is that we can limit the maximum number of attempts in a test item, as it is not influenced by attempt-specific parameters such as γ parameters from later attempts (Tutz, 1990). This is especially useful when a sample size is not large enough to reliably estimate attempt-specific parameters for later attempts. In addition, thanks to the future-agnostic property of SIRT models, we can also reuse the same item parameters and collapse certain categories when only a smaller number of attempts is allowed.

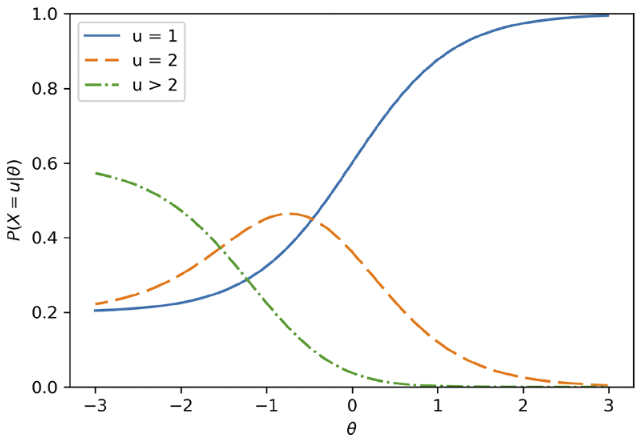


Figure 4. Item category response function when the maximum number of attempts is 3: $a = 1.7, b = 0.0, \gamma_2 = 1$, and $K = 5$.

Table 2. Family of SIRT-MM models

Constraint	No. of parameters	Description
a_{ju}, b_j, and γ_{ju} are unconstrained	2M(K-1)	The SIRT-MM model with the highest degrees of freedom
$a_{ju} = a_u$	(M + 1)(K-1)	
$a_{ju} = a_j$	MK	The second SIRT-MM model we formulated (Eq. (22))
$a_{ju} = a_j, \gamma_{ju} = 0$ for all $u = 3, \ldots, K-1$	3M	A reduced version of the second SIRT-MM model
$a_{ju} = a_j, \gamma_{ju} = 0$ for all $u = 2, \ldots, K-1$	2M	The simplest SIRT-MM model we formulated (Eq. (14))*
⋮		
$a_j = 1, \gamma_{ju} = 0$	M	The simplest SIRT-MM model with fixed a_j parameters*

Note: The second column shows the number of item parameters where M is the number of items and K is the number of answer choices. The models with * at the end of description have the homogeneity of distractors and attempt-invariance assumptions.

For example, Figure 4 shows the item category response functions used in Figure 2a when we set the maximum number of attempts to three. Simply, these are the item category response functions shown in Figure 2a, but $P(X = 3|\theta)$, $P(X = 4|\theta)$, and $P(X = 5|\theta)$ are collapsed into one category. In this example, only γ_2 is relevant, and no matter what true value of γ_3 or γ_4 would be, the SIRT-MM model yields exactly the same model when the maximum number of attempts is three.

2.6. Summary of different parameterizations of SIRT-MM models

Bergner et al. (2019)) summarized existing SIRT models in a table. Similarly, we summarize in Table 2 a family of SIRT-MM models with different constraints. We denote the subject j for individual items and u for the number of attempts. The number of parameters depends on constraints imposed or lifted. The basic SIRT-MM model introduced first in this article is actually a constrained version of the more general SIRT-MM models when $\gamma_{ju} \equiv 0$. We recommend that a model should be selected based on the sample size and model fit statistics such as likelihood ratio tests, Akaike information criterion (AIC; Akaike, 1973) and Bayesian information criterion (BIC; Schwarz, 1978). Later we evaluate the accuracy of model selection using AIC and BIC in the simulation study section.

2.7. Item parameter estimation of SIRT-MM

In this article, we implement marginal maximum likelihood estimation (MMLE) for estimating the item parameters for SIRT-MM models (Bock & Aitkin, 1981). Once the item parameters are estimated, we will estimate θ after treating estimated item parameters as fixed.

In MMLE for item parameters, the likelihood function,

$$L = \int_{-\infty}^{\infty} \prod_{i=1}^{N} \prod_{j=1}^{M} P(X_{ij} = u_{ij}|\theta)\Phi(\theta)d\theta \tag{23}$$

is maximized, where X_{ij} is a random variable representing the number of attempts examinee i needed to submit a correct answer on item j, u_{ij} is the actual number of attempts taken by examinee i to answer item j correctly, N is the number of examinees, M is the number of items, and $\Phi(\theta)$ is a probability density function for the population. Typically, the standard normal distribution is used for $\Phi(\theta)$.

To maximize the likelihood function, we use the log-likelihood, denoted as $\ln L$, instead. Consequently, we require the gradient and Hessian of $\ln P(X_{ij} = u_{ij}|\theta)$ with respect to a parameter of interest to apply Newton's method for maximizing the likelihood function. However, computing the value of the log-likelihood function is not straightforward because the equation contains an integral. In practice, an EM algorithm that uses Gauss–Hermite quadratures is used to compute the marginal likelihood. One should refer to the works by Bock and Aitkin (1981); Muraki (1992) for the details of implementing an EM algorithm for parameter estimation. We will suppress the subscript i next for simplicity.

The generic solutions of the gradient and Hessian of $\ln P(X_j = u|\theta)$ are rather straightforward. Suppose $\phi \in \{a, b, \gamma\}$ and $\omega \in \{a, b, \gamma\}$ are the parameters of interest:

$$\begin{aligned}
\frac{\partial}{\partial \phi} \ln P(X_j = u|\theta) &= \frac{\partial z_{ju}}{\partial \phi} A_{ju} - \sum_{k=1}^{u} \frac{\partial z_{ju}}{\partial \phi} B_{ju}; \\
\frac{\partial^2}{\partial \phi \partial \omega} \ln P(X_j = u|\theta) &= \frac{\partial^2 z_{ju}}{\partial \phi \partial \omega} A_{ju} + \frac{\partial z_{ju}}{\partial \phi} \frac{\partial z_{ju}}{\partial \omega} (A_{ju} - A_{ju}^2) \\
&\quad - \sum_{k=1}^{u} \left(\frac{\partial^2 z_{ju}}{\partial \phi \partial \omega} B_{ju} + \frac{\partial z_{ju}}{\partial \phi} \frac{\partial z_{ju}}{\partial \omega} (B_{ju} - B_{ju}^2) \right),
\end{aligned} \tag{24}$$

where $z_{ju} = a_j(\theta - b_j + \gamma_{ju}), A_{ju} = \frac{K\exp(z_{ju})}{1+K\exp(z_{ju})}$, and $B_{ju} = \frac{K\exp(z_{ju})}{K-i+1+K\exp(z_{ju})}$. Especially, $-\frac{\partial^2}{\partial\phi\partial\omega} \ln P(X_j = u|\theta)$ is called the observed information function and $-\mathbb{E}[\frac{\partial^2}{\partial\phi\partial\omega} \ln P(X_j = u|\theta)] = -\sum_{u=1}^{K} P(X_j = u|\theta) \frac{\partial^2}{\partial\phi\partial\omega} \ln P(X_j = u|\theta)$ is called the expected or Fisher information function of an item. In addition, the Fisher information function of an item is often simply referred to as an item information function. When we estimate a simple model with fewer parameters by setting $\gamma_{ju} = 0$ for any u, we only need to set these values to zero in $z_{ju} = a_j(\theta - b_j + \gamma_{ju})$ and use the same equations, Eq. (24).

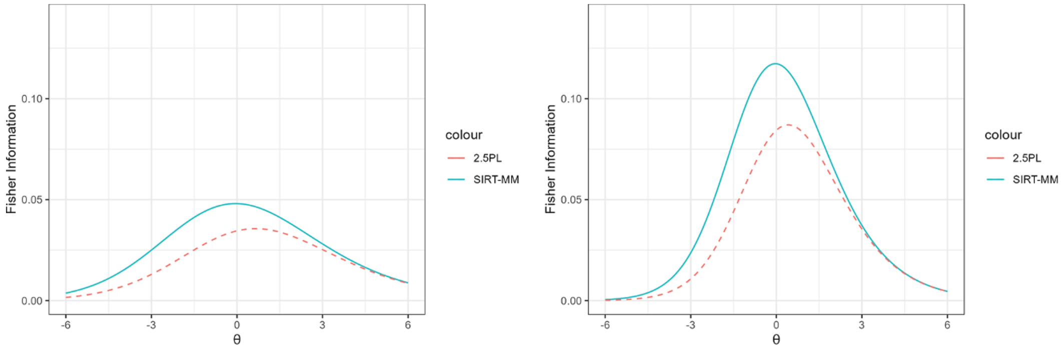
2.8. Person parameter estimation

There exist three popular approaches for estimating person parameters: maximum likelihood estimation (MLE), maximum A posteriori (MAP), and expected A posteriori (EAP) (De Ayala, 2009). MLE maximizes the log-likelihood of a response pattern by Newton's method, MAP uses the mode of the posterior distribution of an θ estimate (typically the standard normal distribution is used for prior), and EAP uses the mean of the posterior distribution of an θ estimate (De Ayala, 2009). In our model, MLE could be obtained by maximizing

$$\ln L_{Resp} = \sum_{j=1}^{M} \ln P(X_{ij} = u_{ij}|\theta_i) \tag{25}$$

with respect to θ_i where θ_i is the latent ability of examinee i, which could be done by Newton's method using Eq. (25). One issue in using MLE is that it cannot provide a θ estimate when a response pattern is all 1 or K. Also, it is known that the mean squared error of θ estimates by EAP is smaller than that obtained by using MLE although its estimation bias is increased (De Ayala, 2009; Lord, 1986). Thus, we use EAP in our simulation study.

https://doi.org/10.1017/psy.2024.18 Published online by Cambridge University Press



(a) $a = 0.48$

(b) $a = 0.75$

Figure 5. Fisher information of SIRT-MM models with $K = 4$, $b = 0$, and $\gamma_u = 0$ for $u = 2$ and 3, and different a; and the corresponding 2.5PL models with $\gamma_1 \equiv 0$.

2.9. Fisher information and standard errors

Under regularity conditions, the Fisher information of item parameter $\phi \in \{a, b, \gamma\}$ is $-\mathbb{E}[\frac{\partial^2}{\partial \phi^2} \ln L]$ and that of θ is $-\mathbb{E}[\frac{\partial^2}{\partial \theta^2} \ln L_{Resp}]$. Thus, we can calculate the standard errors of estimates in $\{\theta, a, b, \gamma\}$, which is inversely related to the square root of the corresponding Fisher information. Thus, the standard error of item parameter ϕ is

$$SE_\phi = \frac{1}{\sqrt{-\mathbb{E}[\frac{\partial^2}{\partial \phi^2} \ln L]}}. \tag{26}$$

Similarly, the standard error of measurement (SEM), which is the standard error of θ is

$$SEM = \frac{1}{\sqrt{-\mathbb{E}[\frac{\partial^2}{\partial \theta^2} \ln L_{Resp}]}}. \tag{27}$$

However, the SEM as defined in the above formula is based on MLE. In this study, since we use EAP, we decide to capture the variation in the person parameter estimates using the empirical SE instead.

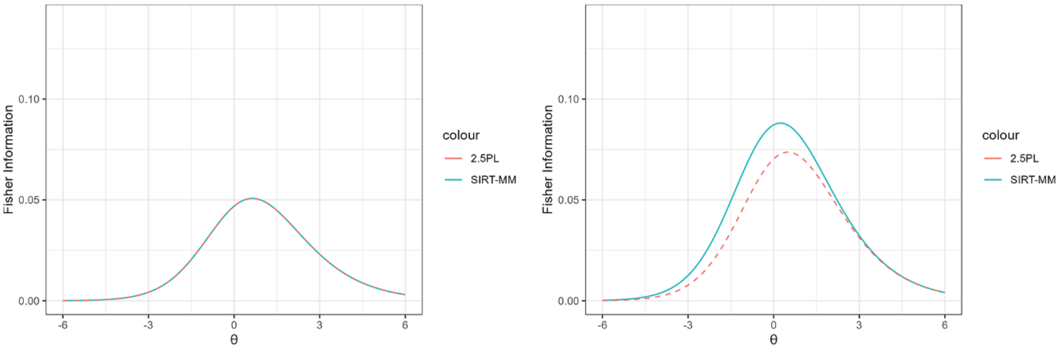
2.10. Item information

Item information is Fisher information computed with respect to θ for any single item, which is a measure of how much an item contributes to reducing the uncertainty about θ estimates (De Ayala, 2009). We can compare item information computed by using SIRT-MM models (which captures information from multiple attempts) against its corresponding 2.5PL model (i.e., the 3PL model with a fixed guessing parameter of 1/K) to demonstrate how much SIRT-MM models potentially improve the accuracy of θ estimates. For example,

Figure 5 shows the item information of SIRT-MM models with $b = 0, K = 4$, and $\gamma_u = 0$ for $u = 2$ and 3, two levels of a parameters ($a = 0.482$ in the left panel and and $a = 0.75$ in the right panel), and its corresponding 2.5PL model. As with the 2.5PL model, SIRT-MM models provide more Fisher information as the a parameter increases. It is noteworthy that for lower θ, SIRT-MM models can yield more information than their 2.5PL counterparts. This is because though reattempts we can gain more information about examinees who fail the first attempt, which is more likely to happen for examinees with lower θ. Conversely, for higher θ, both models have similar information because examinees with higher θ typically only need one attempt to reach the correct answer.

Figure 6 shows the item information of SIRT-MM models with $a = 0.75, b = 0$, and $\gamma_u = 0$ for $u =$ 2 and 3, two levels of K parameters ($K = 2$ in the left panel and and $K = 3$ in the right panel), and its

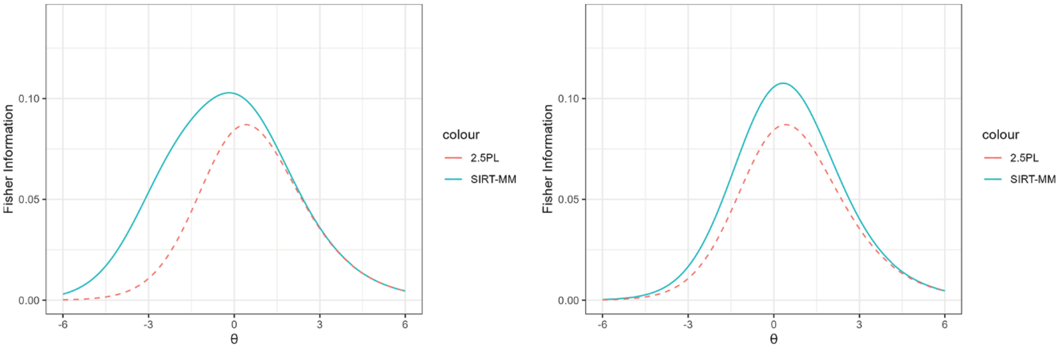
https://doi.org/10.1017/psy.2024.18 Published online by Cambridge University Press



(a) $K = 2$

(b) $K = 3$

Figure 6. Fisher information of SIRT-MM models with $a = 0.75$, $b = 0$, $\gamma_u = 0$ for $u = 2$ and 3, and different K; and the corresponding 2.5PL models with $\gamma_1 \equiv 0$.



(a) $\gamma_2 = 2$

(b) $\gamma_2 = -2$

Figure 7. Fisher information of SIRT-MM models with $a = 0.75$, $b = 0$, $\gamma_3 = 0$, $K = 4$, and different γ_2; and the corresponding 2.5PL models with $\gamma_1 \equiv 0$.

corresponding 2.5PL model. The left panel demonstrates that the 2.5PL models can be considered a special case of SIRT-MM models when $K = 2$. Also, the comparison between the two panels shows that the item information increases for both 2.5PL and SIRT-MM models when K increases because the chance of guessing is reduced.

Figure 7 shows the item information of SIRT-MM models with $a = 0.75$, $b = 0$, and $\gamma_3 = 0$ for $u = 2$ and 3, $K = 4$, two levels of γ_2 ($\gamma_2 = 2$ on the left, and $\gamma_2 = -2$ on the right), and the corresponding 2.5PL models. Changing the γ_2 will affect the amount of item information in lower θ. When γ_2 is positive, we can gain more item information in lower θ than when it is negative because in the latter case examinees with lower θ would not be able to differentiate among distractors and behave similarly to random guessing after the first attempt, as in the example of factual knowledge item.

3. Simulation studies

3.1. Simulation design

We conducted three simulation studies on: model selection, item parameter recovery, and person parameter recovery, respectively. A high-level description of the simulation design shared by these simulation studies is provided here. First, we generated response matrices from the SIRT-MM models. Second, item parameters were estimated by MMLE using an EM algorithm (Bock & Aitkin, 1981)

https://doi.org/10.1017/psy.2024.18 Published online by Cambridge University Press

implemented in R and C++. We provide the R package on GitHub https://github.com/luyikei/sirtmm to fit the SIRT-MM models. A standard normal prior was used for θ in MMLE. Third, with the estimated item parameters considered fixed, person parameters were estimated by EAP. Here, a standard normal prior was used for θ again. Generally, our simulation design followed Reise and Yu (1990).

The first simulation evaluated model selection using AIC and BIC to identify the best model, among SIRT-MM models with all combinations of freely estimated γ parameters to fit multiple attempt data. In addition, we also compared SIRT-MM models in terms of model fit against Graded response model (GRM; Samejima, 1969), since previous research showed that GRM could also be used for AUC (Attali, 2011), as well as Nominal response model (NRM; Bock, 1972). The simulation conditions are specified as $N = 500$ and $4{,}000$, $M = 30$, $K = 4$, $\theta \sim N(0,1)$, $b_j \sim \text{Unif}(-2,2)$, $a_j \sim \text{Unif}(0.75,1.33)$, and $\gamma_{ju} \sim \text{Unif}(-1,1)$. We analyzed the impact of varying sample sizes, $N = 500$ and $N = 4000$, to assess how differences in sample size influence model selection performance. For each replication, we simulated responses from all possible SIRT-MM models and fit all candidate models (SIRT-MM, GRM, and NRM models) to each response matrix. Over 100 replications, we were able to obtain model selection with AIC and BIC.

Then, we conducted simulation studies to evaluate item and person parameter recovery of SIRT-MM models. As item and person parameter estimations take place at different stages, we evaluated them separately. The second simulation study investigates item parameter recovery of SIRT-MM models under different sample sizes and test length conditions. The number of answer choices is $K = 4$. We evaluated all the combinations of (1) the sample size: $N = 250, 500, 1{,}000, 2{,}000, 4{,}000, 8{,}000, 16{,}000$; (2) the number of items: $M = 15, 20, 25, 30$; and (3) the number of $\gamma_{ju} \sim \text{Unif}(-1,1)$. We also evaluated more simulation conditions varying other factors; however, they are not fully crossed with conditions (1)–(3), as it would result in an unrealistic number of simulation conditions. Specifically, we evaluated conditions varying (4) θ distribution: normal ($N(0,1)$), uniform ($\text{Unif}(-3,3)$), and skewed normal distribution $sn(\xi = -1.5, \omega = 2, \alpha = 6)$ using sn package (Azzalini, 2022); (5) item discrimination parameter, α_j: sampled from low ($\text{Unif}(0.44,0.75)$), middle ($\text{Unif}(0.58,0.98)$), high ($\text{Unif}(0.75,1.33)$), and all ($\text{Unif}(0.44,1.33)$) ranges; (6) item difficulty parameter: sampled from all ($\text{Unif}(-2,2)$) and high ($\text{Unif}(0,2)$) ranges; and (7) setting the maximum number of attempts of 2 vs. 4. The simulation conditions in (4)–(7) are evaluated with different sample sizes but share the same baseline condition, which is specified as follows: $M = 20$, $a_j \sim \text{Unif}(0.75,1.33)$, $b_j \sim \text{Unif}(-2,2)$, for γ_{ju} parameters, only $\gamma_{j2} \sim \text{Unif}(-1,1)$ is specified as a freely estimated parameter, and the θ distribution is the standard normal distribution ($\theta \sim N(0,1)$). The skewed distribution for θ is positively skewed in order to evaluate the performance of person parameter recovery for a low-ability population, for which the SIRT-MM models are good candidates. For the same reason, the item difficulty parameter has a condition where only relatively difficult items exist. The convergence rate of item parameter estimation was reported for each simulation condition. Standard errors (SE) for the item parameters, bias, and root mean square error (RMSE) were used as primary indices to examine the quality of parameter estimates, which were obtained for the converged conditions. Out of 100 replications, we calculated the averages of metrics from all converged replications across all conditions.

The third simulation study evaluated person parameter recovery following the same baseline condition as the first simulation: $M = 20$, $a_j \sim \text{Unif}(0.75,1.33)$, $b_j \sim \text{Unif}(-2,2)$. For γ_{ju} parameters, only $\gamma_{j2} \sim \text{Unif}(-1,1)$ is specified as a freely estimated parameter, and the θ distribution is the standard normal distribution ($\theta \sim N(0,1)$). θ was estimated by EAP treating item estimates as fixed. To evaluate person parameter recovery, we used the 2.5PL model as a baseline for comparison since multiple-attempt responses could be converted to 0/1 scoring if we only take first-attempt data. Note that model fit and selection are evaluated in a different simulation study and the purpose of this comparison is to show how much improvement in person parameter recovery could be gained by just allowing multiple attempts using the same test items. We used mirt package for estimating the 2.5PL model (Chalmers, 2012). In addition to bias and RMSE, the Pearson correlation coefficient was also used to assess the recovery accuracy for θ. When we evaluate correlation, we also included the results of

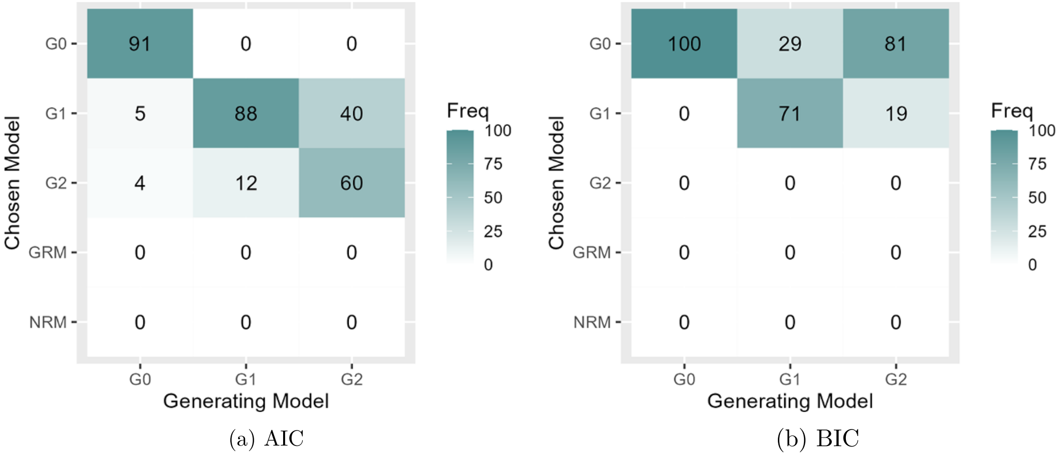


Figure 8. Model selection performance of AIC and BIC for SIRT-MM models when data are generated from SIRT-MM models with $N = 500$, $M = 30, K = 4, \theta \sim N(0, 1), b_j \sim \text{Unif}(-2, 2), a_j \sim \text{Unif}(0.75, 1.33)$, and $\gamma_{ju} \sim \text{Unif}(-1, 1)$. The freely estimated γ_{ju} are denoted as *Ga* where *a* is the number of γ_{ju} parameters for all *u*.

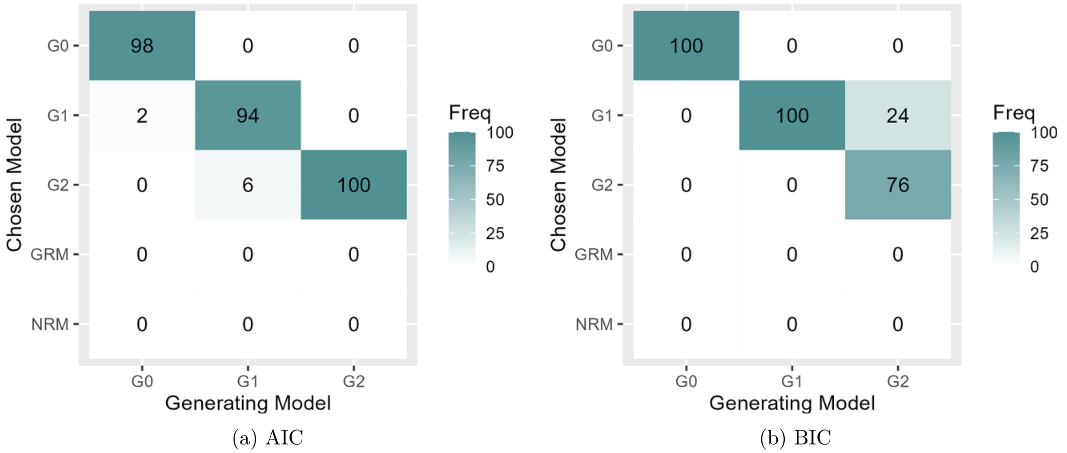
a popular scoring scheme in CTT which grants full credit for the successful first attempt to an item, half credit for the successful second attempt, one-tenth credit for the successful third attempt, and zero credit otherwise (Slepkov & Godfrey, 2019). Results were presented by taking the averages of metrics calculated from 100 replications for all conditions.

3.2. Results

3.3. Model selection

Figures 8 and 9 present the model selection performance of AIC and BIC for SIRT-MM models with $N = 500$ and $N = 4{,}000$. First, both AIC and BIC successfully select an SIRT-MM model over GRM or NRM when responses are simulated from an SIRT-MM model. Second, for selecting the correct SIRT-MM model from all the variants of SIRT-MM models, AIC selects the correct model the majority of times regardless of N, and both AIC and BIC perform well with larger N. Specifically, when $N = 500$, AIC could identify the correct model about 90% of the time and BIC could identify the correct model about 86% of the time from models with or without γ_{j2}. However, for the data generating model incorporating both γ_{j2} and γ_{j3}, a sample size of N = 500 results in AIC correctly identifying the model 60% of the time, while BIC never identifies the correct model. When $N = 4{,}000$, AIC could identify the correct model about 97% of the time and BIC could identify the correct model about 92% of the time from all the models. Between the two, AIC seems to outperform BIC, as BIC could under-specify the model, though in a small number of cases AIC could over-specify the model. Specifically, with $N = 500$, in 21 out of a total of 300 generated cases across the conditions AIC over-specified the model while BIC under-specifies the model with γ_{j2} in 29 out of 100 cases and consistently under-specifies the model with both γ_{j2} and γ_{j3} (i.e., 100 out of 100 cases). In contrast, AIC under-specifies the model containing both γ_{j2} and γ_{j3} in 40 out of 100 cases. When $N = 4{,}000$, AIC identifies γ_{j2} parameters when no γ_{ju} was included in the generating model in 2 out of 100 cases, and AIC identifies γ_{j2} and γ_{j3} when the generating model only had γ_{j2} parameter in 6 out of 100 cases. On the other hand, it is worth noting that there are 24 out of all the generated cases (300) across the conditions where BIC identifies only γ_{j2} parameters while the data generating model includes both γ_{j2} and γ_{j3}. In our simulation results under both sample size conditions, AIC is more accurate in selecting the true model and BIC in some cases picks an over-simplified model.

https://doi.org/10.1017/psy.2024.18 Published online by Cambridge University Press



(a) AIC (b) BIC

Figure 9. Model selection performance of AIC and BIC for SIRT-MM models when data are generated from SIRT-MM models with $N = 4{,}000$, $M = 30$, $K = 4$, $\theta \sim N(0,1)$, $b_j \sim \text{Unif}(-2,2)$, $a_j \sim \text{Unif}(0.75,1.33)$, and $\gamma_{ju} \sim \text{Unif}(-1,1)$. The freely estimated γ_{ju} are denoted as Ga where a is the number of γ_{ju} parameters for all u.

3.4. Item parameter recovery

Tables 3–5 present the item parameter recovery statistics varying sample size, number of items, and number of effective γ parameters when $\theta \sim N(0,1)$, $a_j \sim \text{Unif}(0.75,1.33)$, $b_j \sim \text{Unif}(-2,2)$. The results show that, as N gets larger, the SE and RMSE for the item parameter estimates decrease and the bias quickly converges to zero in all the conditions, suggesting that our estimation method could yield satisfactory item parameter recovery for all conditions given a large enough N. Although M has a smaller effect on item recovery statistics compared to N, generally larger M also leads to better item parameter estimates.

Different SIRT-MM models tend to vary in their sample size requirements, and thus investigated separately. Table 3 presents item parameter recovery results when all γ_{ju} parameters are constrained to be zero. This model has the fewest number of item parameters among all variants, as only a_j and b_j are estimated. Please note that this is not the same as a 2.5PL model, because we simulated a maximum of four attempts. Item parameter estimation converged in all conditions, except for one single case when $N = 250$ and $M = 20$. The RMSE for the item parameter estimates are smaller than 0.6 in all N and M conditions, smaller than 0.3 with $N \geq 500$, and could be smaller than 0.1 with $N \geq 4{,}000$. Overall, these results show that the item parameters from the simplest SIRT-MM model can be recovered very well when the model fits the data with a reasonable sample size (e.g., $N = 500$ or more).

Table 4 presents item recovery statistics when only $\gamma_{j2} \sim \text{Unif}(-1,1)$ is specified as a freely estimated parameter. Item parameter estimation generally converged in all conditions although when $N = 250$, there is a 2%–5% chance of non-convergence. The RMSE for the item parameter estimates are smaller than 0.3 when $N \geq 1{,}000$. However, if we can compromise the accuracy of γ_{ju} a little bit, $N \geq 500$ is also acceptable since the RMSE for γ_{ju} will not be larger than 0.5 when $N = 500$. Table 5 presents item recovery statistics when $\gamma_{j2}, \gamma_{j3} \sim \text{Unif}(-1,1)$ are specified as freely estimated parameters. We do not recommend $N <= 1{,}000$ for estimating both γ_{j2} and γ_{j3} because the RMSE for γ_{j3} are generally very high and the convergence rates could be low. On the other hand, the RMSE for all the item parameters are smaller than 0.35 when $N \geq 4000$.

In Supplementary Material, we include additional tables, Supplementary Tables S1–S4, which present the item parameter recovery statistics varying θ distributions, item discrimination parameters, item difficulty parameters and the maximum number of attempts respectively. Especially, Supplementary Table S1 shows that an SIRT-MM models works better with a positively skewed θ distribution than the 2.5PL model, and Supplementary Table S3 shows that having a relatively difficult test (by keeping b_j

https://doi.org/10.1017/psy.2024.18 Published online by Cambridge University Press

Table 3. Item recovery statistics for items without γ_{ju}

		SE		BIAS		RMSE		CONV
M	*N*	b_j	a_j	b_j	a_j	b_j	a_j	
15	250	0.30	0.23	0.02	0.02	0.40	0.32	1.00
	500	0.18	0.16	−0.00	0.01	0.23	0.21	1.00
	1,000	0.12	0.11	0.01	0.01	0.16	0.15	1.00
	2,000	0.09	0.08	0.01	−0.00	0.12	0.11	1.00
	4,000	0.06	0.06	0.00	0.00	0.08	0.07	1.00
	8,000	0.04	0.04	0.01	−0.00	0.06	0.05	1.00
	16,000	0.03	0.03	0.01	−0.00	0.04	0.04	1.00
20	250	0.30	0.23	−0.00	0.03	0.40	0.31	0.99
	500	0.19	0.16	−0.01	0.00	0.25	0.20	1.00
	1,000	0.12	0.11	0.01	0.00	0.16	0.15	1.00
	2,000	0.09	0.08	−0.00	−0.00	0.11	0.10	1.00
	4,000	0.06	0.06	0.00	−0.00	0.08	0.07	1.00
	8,000	0.04	0.04	0.00	−0.00	0.05	0.05	1.00
	16,000	0.03	0.03	0.01	−0.01	0.04	0.04	1.00
25	250	0.28	0.23	−0.00	0.02	0.36	0.29	1.00
	500	0.18	0.16	−0.01	0.01	0.23	0.20	1.00
	1,000	0.12	0.11	−0.00	0.00	0.15	0.14	1.00
	2,000	0.09	0.08	−0.00	−0.00	0.10	0.09	1.00
	4,000	0.06	0.06	0.01	−0.00	0.07	0.07	1.00
	8,000	0.04	0.04	0.00	−0.01	0.05	0.05	1.00
	16,000	0.03	0.03	0.00	−0.01	0.04	0.03	1.00
30	250	0.28	0.23	−0.00	0.02	0.36	0.29	1.00
	500	0.18	0.16	−0.01	0.01	0.23	0.20	1.00
	1,000	0.12	0.11	−0.00	0.00	0.16	0.14	1.00
	2,000	0.09	0.08	−0.00	−0.00	0.10	0.10	1.00
	4,000	0.06	0.06	−0.00	−0.01	0.07	0.07	1.00
	8,000	0.04	0.04	−0.00	−0.01	0.05	0.05	1.00
	16,000	0.03	0.03	−0.00	−0.01	0.04	0.03	1.00

Note: “CONV” stands for convergence rate for a simulation condition.

parameters to a high range) does not seem to affect item parameter estimation much for SIRT-MM models. This is because SIRT-MM models can glean more item information in the lower θ range from multiple attempts. Please refer to the Supplementary Material for further elaboration.

In sum, although all the factors more or less affect the accuracy of item parameter estimates, having a reasonably large sample size enables quality item parameter estimates. For a simple SIRT-MM model, we recommend $N = 500$ or more. For more complex SIRT-MM models, $N = 1{,}000$ or 2,000 or more might be needed.

Table 4. Item recovery statistics for items with γ_{j2}

		SE			BIAS			RMSE			CONV
M	N	b_j	a_j	γ_{j2}	b_j	a_j	γ_{j2}	b_j	a_j	γ_{j2}	
15	250	0.37	0.23	0.55	0.03	0.03	0.03	0.52	0.33	1.00	0.98
	500	0.20	0.16	0.35	0.01	0.03	0.01	0.26	0.24	0.49	1.00
	1,000	0.13	0.11	0.24	0.01	0.01	−0.01	0.17	0.15	0.27	1.00
	2,000	0.09	0.08	0.17	0.01	0.00	0.00	0.12	0.11	0.18	1.00
	4,000	0.06	0.06	0.11	0.01	−0.00	0.00	0.08	0.08	0.12	1.00
	8,000	0.05	0.04	0.08	0.01	−0.00	0.00	0.06	0.05	0.09	1.00
	16,000	0.03	0.03	0.06	0.01	−0.00	−0.00	0.04	0.04	0.06	1.00
20	250	0.30	0.24	0.51	0.01	0.05	0.06	0.39	0.32	0.78	0.95
	500	0.20	0.16	0.34	0.01	0.02	−0.00	0.25	0.22	0.36	1.00
	1,000	0.13	0.11	0.24	0.00	0.00	−0.01	0.17	0.15	0.27	1.00
	2,000	0.09	0.08	0.17	0.01	−0.00	−0.00	0.12	0.10	0.18	1.00
	4,000	0.06	0.06	0.12	0.01	−0.00	−0.00	0.08	0.07	0.13	1.00
	8,000	0.05	0.04	0.08	0.00	−0.00	−0.00	0.06	0.05	0.09	1.00
	16,000	0.03	0.03	0.06	0.00	−0.00	0.00	0.04	0.04	0.06	1.00
25	250	0.30	0.24	0.52	0.01	0.03	0.04	0.37	0.31	0.77	0.97
	500	0.20	0.16	0.35	−0.00	0.01	0.01	0.24	0.21	0.43	1.00
	1,000	0.13	0.11	0.24	0.00	0.00	−0.00	0.16	0.14	0.25	1.00
	2,000	0.09	0.08	0.17	0.01	−0.00	−0.00	0.11	0.10	0.18	1.00
	4,000	0.06	0.06	0.12	0.00	−0.00	−0.00	0.08	0.07	0.12	1.00
	8,000	0.05	0.04	0.08	0.00	−0.01	−0.00	0.06	0.05	0.09	1.00
	16,000	0.03	0.03	0.06	0.00	−0.01	−0.00	0.04	0.04	0.06	1.00
30	250	0.30	0.23	0.58	0.00	0.02	0.07	0.38	0.30	1.11	0.97
	500	0.20	0.16	0.35	0.00	0.01	0.00	0.24	0.20	0.38	1.00
	1,000	0.13	0.11	0.24	−0.00	−0.00	0.01	0.16	0.14	0.25	1.00
	2,000	0.09	0.08	0.17	0.01	−0.00	0.00	0.11	0.10	0.18	1.00
	4,000	0.07	0.06	0.12	0.00	−0.01	−0.00	0.08	0.07	0.12	1.00
	8,000	0.05	0.04	0.08	−0.00	−0.01	−0.00	0.05	0.05	0.09	1.00
	16,000	0.03	0.03	0.06	0.00	−0.01	−0.00	0.04	0.03	0.06	1.00

Note: "CONV" stands for convergence rate for a simulation condition.

3.5. *Person parameter recovery*

Figure 10 shows the person recovery statistics varying the number of items, M, when $\theta \sim N(0,1)$, $a \sim \text{Unif}(0.75, 1.33)$, $b \sim \text{Unif}(-2,2)$, and $\gamma_{j2} \sim \text{Unif}(-1,1)$. The left panel shows RMSE, the middle panel shows bias and the right panel shows correlations between the true and estimated θ. The RMSE for θ estimated by the SIRT-MM model with freely estimated γ_{ju} parameters is consistently smaller than those estimated by the 2.5PL model. The RMSE of θ estimates is quite low even with $N = 250$, indicating that the person parameter estimation can be robust even at very small sample sizes. The bias for θ estimated both by the SIRT-MM model and the 2.5 PL model are consistently close to zero. The

Table 5. Item recovery statistics for an item with γ_{j2} and γ_{j3}

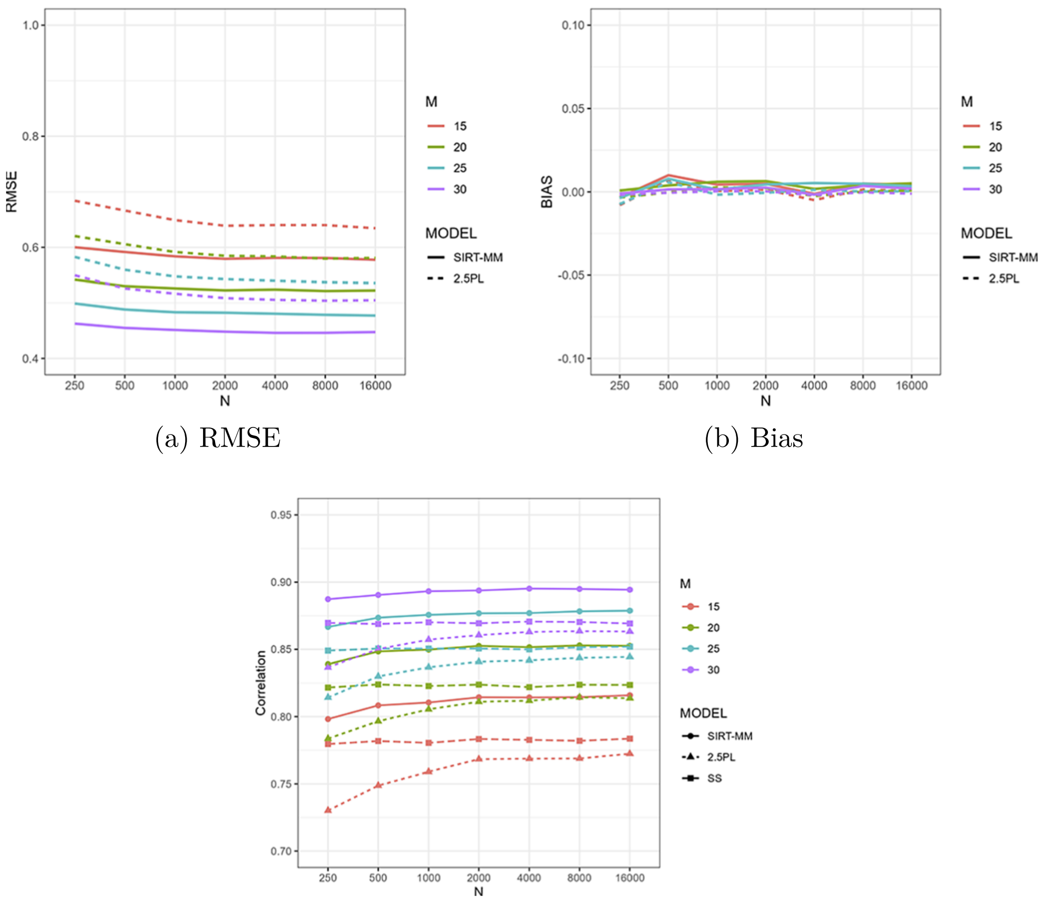
		SE				BIAS				RMSE				CONV
M	*N*	b_j	a_j	γ_{j2}	γ_{j3}	b_j	a_j	γ_{j2}	γ_{j3}	b_j	a_j	γ_{j2}	γ_{j3}	
15	250	0.30	0.24	0.51	1.03	0.04	0.07	−0.03	1.97	0.39	0.34	0.92	7.30	0.67
	500	0.19	0.17	0.34	0.67	0.02	0.03	−0.00	0.94	0.24	0.24	0.49	4.32	0.89
	1,000	0.14	0.12	0.24	0.49	0.01	0.01	−0.00	0.43	0.18	0.16	0.27	2.02	0.97
	2,000	0.10	0.08	0.17	0.34	0.01	−0.00	−0.00	0.07	0.13	0.12	0.19	0.62	1.00
	4,000	0.07	0.06	0.12	0.25	0.01	−0.01	−0.01	−0.00	0.09	0.08	0.13	0.29	1.00
	8,000	0.05	0.04	0.08	0.17	0.01	−0.00	0.00	0.00	0.06	0.06	0.09	0.19	1.00
	16,000	0.03	0.03	0.06	0.12	0.01	−0.00	0.00	0.01	0.04	0.04	0.06	0.14	1.00
20	250	0.32	0.24	0.53	1.05	0.03	0.06	−0.01	2.05	0.44	0.36	1.01	8.57	0.53
	500	0.20	0.17	0.35	0.70	0.02	0.01	−0.02	0.91	0.26	0.23	0.38	5.01	0.89
	1,000	0.14	0.12	0.24	0.49	0.01	0.01	0.00	0.48	0.17	0.16	0.26	2.57	0.99
	2,000	0.10	0.08	0.17	0.35	−0.00	0.00	−0.01	0.09	0.12	0.11	0.18	0.78	0.99
	4,000	0.07	0.06	0.12	0.24	0.01	0.00	0.00	0.01	0.08	0.08	0.13	0.33	1.00
	8,000	0.05	0.04	0.08	0.17	0.00	−0.00	−0.00	0.00	0.06	0.05	0.09	0.18	1.00
	16,000	0.03	0.03	0.06	0.12	0.00	−0.00	−0.00	0.00	0.04	0.04	0.06	0.13	1.00

(Continued)

Table 5. Continued

		SE				BIAS				RMSE				CONV
M	N	b_j	a_j	γ_{j2}	γ_{j3}	b_j	a_j	γ_{j2}	γ_{j3}	b_j	a_j	γ_{j2}	γ_{j3}	
25	250	0.31	0.24	0.53	0.97	0.03	0.05	0.08	2.39	0.39	0.33	1.16	8.96	0.51
	500	0.20	0.17	0.35	0.69	0.01	0.03	0.03	1.18	0.25	0.22	0.47	5.34	0.90
	1,000	0.14	0.12	0.24	0.49	0.00	0.00	0.01	0.39	0.16	0.15	0.26	2.12	0.97
	2,000	0.10	0.08	0.17	0.34	0.01	0.00	−0.00	0.01	0.12	0.10	0.18	0.39	0.99
	4,000	0.07	0.06	0.12	0.25	0.00	−0.00	−0.00	−0.00	0.08	0.07	0.12	0.28	0.99
	8,000	0.05	0.04	0.08	0.17	0.00	−0.01	0.00	0.00	0.06	0.05	0.09	0.19	1.00
	16,000	0.03	0.03	0.06	0.12	0.00	−0.01	0.00	0.00	0.04	0.04	0.06	0.13	1.00
30	250	0.29	0.24	0.51	0.97	0.00	0.05	0.05	2.10	0.35	0.31	0.86	8.87	0.44
	500	0.20	0.16	0.36	0.69	−0.00	0.01	0.01	0.97	0.25	0.21	0.43	4.93	0.89
	1,000	0.14	0.12	0.24	0.51	−0.00	−0.00	−0.00	0.31	0.16	0.14	0.26	2.14	0.98
	2,000	0.10	0.08	0.17	0.35	−0.00	0.00	0.00	0.10	0.11	0.10	0.18	0.78	0.99
	4,000	0.07	0.06	0.12	0.24	−0.00	−0.01	−0.00	0.01	0.08	0.07	0.13	0.33	1.00
	8,000	0.05	0.04	0.08	0.17	0.00	−0.01	−0.00	0.01	0.06	0.05	0.09	0.20	1.00
	16,000	0.03	0.03	0.06	0.12	0.00	−0.01	0.00	0.00	0.04	0.04	0.06	0.13	1.00

Note: “CONV” stands for convergence rate for a simulation condition.



(a) RMSE

(b) Bias

(c) Correlation

Figure 10. Person parameter statistics when $\theta \sim N(0,1), a_j \sim \text{Unif}(0.75,1.33), b_j \sim \text{Unif}(-2,2)$, and $\gamma_{j2} \sim \text{Unif}(-1,1)$. M is the number of items administered. The scoring scheme used in classical test theory is denoted as SS in the correlation plot.

correlations between true θ and θ estimated by the SIRT-MM model are consistently the highest among the three scoring mechanisms. Typically the CTT scoring scheme outperforms the 2.5PL model in terms of correlation because it still can recover some partial information from multiple attempts.

Figure 11 shows the conditioned RMSE for θ estimates. The SIRT-MM model leads to lower RMSE, especially at the lower range of θ. Thus, the SIRT-MM model could be used for improving person parameter estimates, especially at the low end of the θ.

In Supplementary Material, we include additional figures, Supplementary Figures S3–S6, which present the person parameter recovery statistics varying θ distributions, item discrimination parameters, item difficulty parameters and the maximum number of attempts, respectively. Generally, an SIRT-MM model outperforms the 2.5PL model in all conditions especially in RMSE when the SIRT-MM model is the true model. Please refer to the Supplementary Material for detailed explanations.

4. Empirical analysis

We applied the SIRM-MM model to a real dataset collected from both from college students ($N = 167$) and Prolific ($N = 295$) participants. They took multiple-choice, multiple-attempt trivia questions about

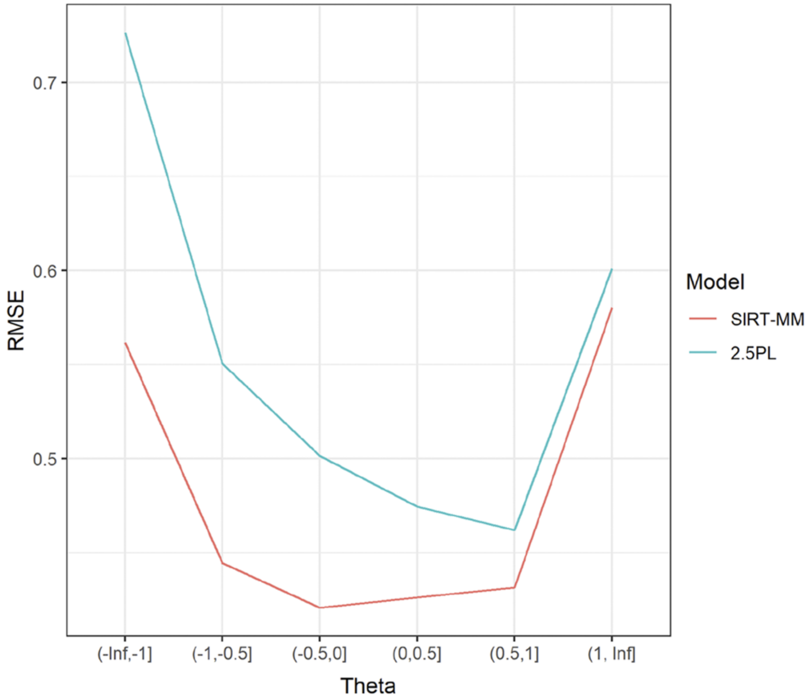


Figure 11. RMSE for θ estimates conditioning on θ when $N = 1,000, M = 25, \theta \sim N(0,1), a_j \sim \text{Unif}(0.75, 1.33), b_j \sim \text{Unif}(-2, 2)$.

Harry Potter through an online platform, following an AUC procedure. The data was collected between May 2023 to March 2024. There was no missing or incomplete responses. To ensure the quality of data and item parameter estimates, we omitted four test items that were generally very difficult and had very few correct responses at the first attempt. The resulting response matrix included multiple-attempt responses of 462 examinees to 27 test items with four answer options. Because the sample size ($N = 462$) was limited, based on the sample size guidelines from our simulation studies, only two candidate SIRT-MM models were fit to the data: (1) the simplest SIRT-MM model without any freely estimated γ_{ju}, and (2) an SIRT-MM model with freely estimated γ_{j2} only. We chose a better model with smaller AIC and BIC values between the two candidate models. We also fitted GRM and NRM for comparison. For this analysis, we fixed the maximum number of attempts to two so only data for the first two attempts were used to fit the two candidate models. We compared the θ estimates and test information function derived using two attempts against those derived from only the first attempt data. In the Supplementary Material, we present the resulting item parameter estimates (Supplementary Table S5), the histograms of the number of attempts for each item (Supplementary Figure S7) and the item category response functions (Supplementary Figure S8).

Table 6 shows the model fit statistics for the two candidate SIRT-MM models, GRM, and NRM. It shows that both SIRT-MM models lead to smaller AIC and BIC than GRM and NRM in AIC, BIC, and negative log likelihood. AIC is the smallest for the SIRT-MM model with freely estimated γ_{j2} and BIC is the smallest for the simplest SIRT-MM model. As AIC is shown to be more accurate in selecting the correct model in our simulation study when $N = 500$, we selected the SIRT-MM model with freely estimated γ_{j2} for the subsequent analysis.

Figure 12 presents the scatter plot of θ estimated by the SIRT-MM with freely estimated γ_{j2} with one- vs. two-attempt data. The result shows that all the θ estimates are generally very similar to each other and there is no outlier that yields very different estimates between one or two attempts. However, the precision of these estimates can be different. Figure 13 presents the test information functions

Table 6. Model-fit statistics for an SIRT-MM model without γ_{ju} and an SIRT-MM model with a freely estimated γ_{j2}

Model	AIC	BIC	Negative log likelihood	Number of parameters
Without γ_{j2}	17,866.36	18,089.68	8,879.18	54
With γ_{j2}	17,778.50	18,113.48	8,808.25	81
GRM	18,896.15	19,231.13	9,367.08	81
NRM	18,832.71	19,279.35	9,308.35	108

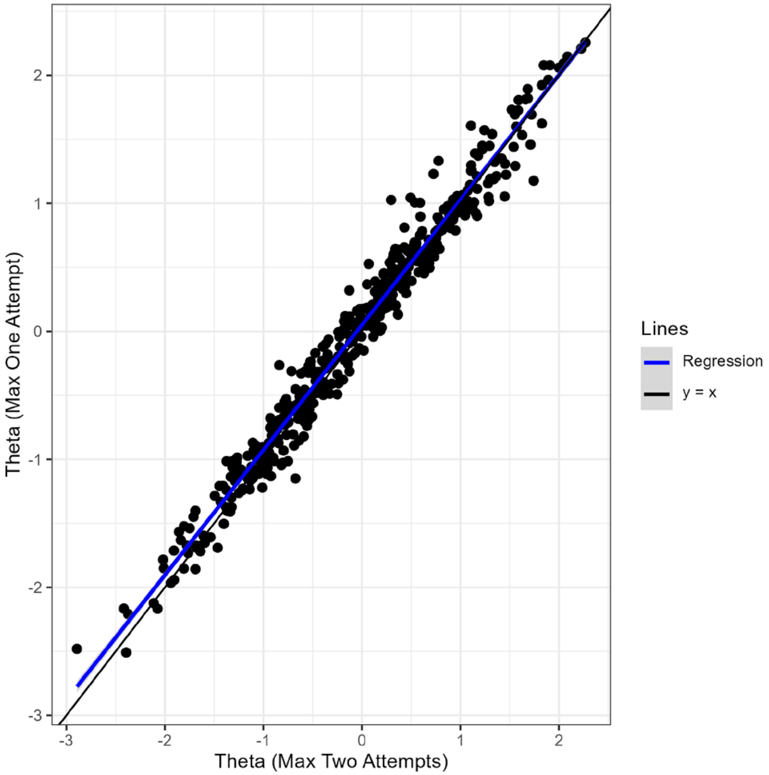


Figure 12. Scatter plot of θ estimated by the SIRT-MM models using only one attempt and two attempts from the real data.

derived from the SIRT-MM with one- vs. two-attempt data. There is a consistent increase of test information from one attempt to two attempts, suggesting that allowing two attempts helps gain more information from examinees, which leads to smaller SE of θ estimates.

5. Discussion

This article has proposed and formally derived a family of new sequential item response models (SIRT-MM models) for multiple-choice, multiple-attempt test items that considers the guessing of multiple-choice test items, and the homogeneity and heterogeneity of distractors. We demonstrated that an SIRT-MM model can be used to glean more information from multiple-choice, multiple-attempt items and to provide better scoring, especially for in the region of smaller θ.

Our simulation study included model selection, and item and person parameter recovery. For model selection, we showed that AIC and BIC never selected GRM or NRM when data were generated from

https://doi.org/10.1017/psy.2024.18 Published online by Cambridge University Press

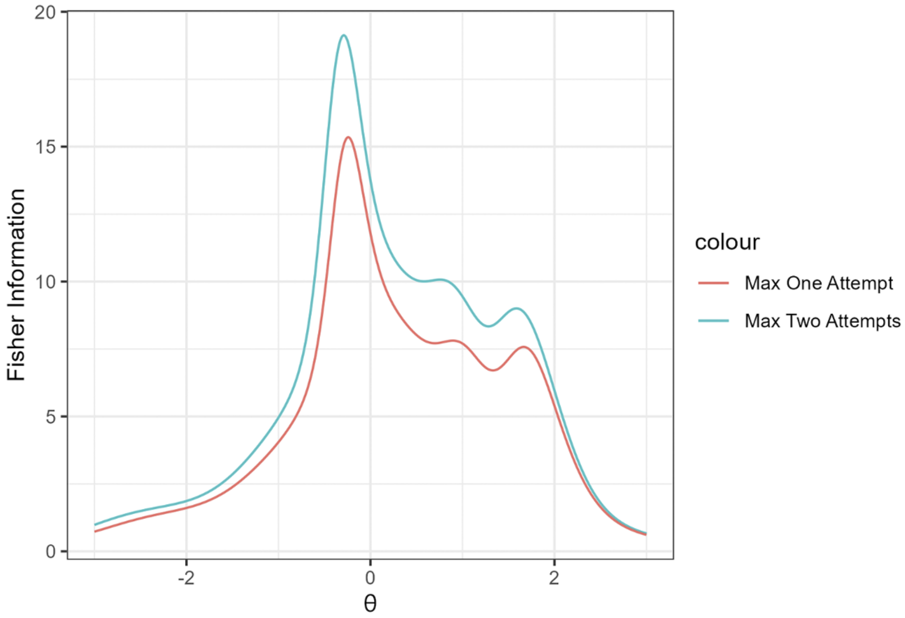


Figure 13. Test information functions of the real data with different maximum numbers of attempts using the item parameter estimates for the SIRT-MM model.

SIRT-MM models, demonstrating the unique utility of the SIRT-MM models to model multiple-attempt data. For item parameter recovery, we showed that our implementation of MMLE could recover item parameters very well with N = 500 for the simplest SIRT-MM model and with N = 1,000 or 2,000 for more complex SIRT-MM models with reasonable test lengths. For person parameter recovery, we showed that an SIRT-MM model consistently outperforms the 2.5PL model in all conditions when former is the true model. Also, the person parameter recovery results suggested that θ can be estimated reasonably well even with a small sample size. Taken together, one could consider adopting a multiple-attempt procedure and SIRT-MM models to improve measurement precision.

One limitation of this study is that we have not fully investigated different possible parameterizations of SIRT-MM models. First, our proposed models do not allow a freely estimated pseudo-guessing parameter for each item. By doing so, SIRT-MM models could be compared against the actual 3PL model, instead of the 2.5PL model. That being said, it is worth noting that a previous study showed that fixing the pseudo-guessing parameter in 3PL model provides a stable and accurate item estimation solution (Han, 2012), and thus our study still provides practical utility. Second, our proposed models do not focus on varying the a_j parameter at each attempt. The concept of allowing the item discrimination parameter to change at each attempt by introducing δ_{ju} parameters is explained and discussed in the Supplementary Material. This extension could be important as item discrimination in a traditional SIRT model could decrease with each attempt (Lyu et al., 2023). Both extensions imply estimating many additional item parameters, which could cause convergence issues or inaccurate item parameters unless we have a huge sample size. Future work could consider regularization for item parameter estimation or Bayesian estimation to help accurately estimate more item parameters even with a smaller sample size.

A few additional limitations should be noted about the current study. First, our simulation did not evaluate all the combinations of simulation conditions for item parameters. The sample size requirement would be different depending on various factors including the complexity of a model and the distributions of true parameters including γ_{ju}. As suggested by a reviewer, interactions between such factors should also be evaluated. In addition, our models did not model learning or growth in this context. For future work, we can consider a use case where growth-curve SIRT-MM models similar to Culpepper (2014) could be formulated and used to track examinees' learning.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psy.2024.18. The R package that includes the estimation program and sample real data are available at https://github.com/luyikei/sirtmm.

Funding statement. Collection of the empirical dataset was made possible by support from the Institute for Scholarship in the Liberal Arts, College of Arts and Letters, University of Notre Dame.

Competing interests. None.

References

Agresti, A. (2013). *Categorical data analysis*. (3rd ed.). John Wiley & Sons Inc.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd international symposium on information theory* (pp. 267–281), Budapest, Hungary.

Attali, Y. (2011). Immediate feedback and opportunity to revise answers: Application of a graded response IRT model. *Applied Psychological Measurement*, *35*(6), 472–479. https://doi.org/10.1177/0146621610381755

Azzalini, A. (2022). The R package sn: The skew-normal and related distributions such as the skew-*t* and the SUN (version 2.0.2). [Computer software manual]. Università degli Studi di Padova, Italia. https://cran.r-project.org/package=sn

Ben-Akiva, M. E. (1985). *Discrete choice analysis: theory and application to travel demand*. MIT Press.

Benson, A. R., Kumar, R., & Tomkins, A. (2016). On the relevance of irrelevant alternatives. In *Proceedings of the 25th international conference on world wide web* (pp. 963–973). Republic and Canton of Geneva, CHE. https://doi.org/10.1145/2872427.2883025

Bergner, Y., Choi, I., & Castellano, K. E. (2019). Item response models for multiple attempts with incomplete data. *Journal of Educational Measurement*, *56*(2), 415–436. https://doi.org/10.1111/jedm.12214

Bizot, E. B., & Goldman, S. H . (1994). The practical impact of irt models and parameters when converting a test to adaptive format. In *Paper presented at the annual meeting of the American educational research association*.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51. https://doi.org/10.1007/BF02291411

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Culpepper, S. A. (2014). If at first you don't succeed, try, try again: Applications of sequential IRT models to cognitive assessments. *Applied Psychological Measurement*, *38*(8), 632–644. https://doi.org/10.1177/0146621614536464

Davis, F. B. (1964). *Educational measurements and their interpretation*. Wadsworth Publishing Company.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

DiBattista, D., Gosse, L., Sinnige-Egger, J.-A., Candale, B., & Sargeson, K. (2009). Grading scheme, test difficulty, and the immediate feedback assessment technique. *The Journal of Experimental Education*, *77*(4), 311–338. https://doi.org/10.3200/JEXE.77.4.311-338

Epstein, M. L., Epstein, B. B., & Brosvic, G. M. (2001). Immediate feedback during academic testing. *Psychological Reports*, *88*(3), 889–894. https://doi.org/10.2466/pr0.2001.88.3.889

Frary, R. B. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. *Applied Psychological Measurement*, *4*(1), 79–90. https://doi.org/10.1177/014662168000400109

Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education*, *2*(1), 79–96. https://doi.org/10.1207/s15324818ame0201_5

Gilman, D. A., & Ferry, P. (1972). Increasing test reliability through self-scoring procedures. *Journal of Educational Measurement*, *9*(3), 205–207. https://doi.org/10.1111/j.1745-3984.1972.tb00953.x

Han, K. T . (2012). Fixing the c parameter in the three-parameter logistic model. *Practical Assessment, Research & Evaluation*, *17*(1), 1–24. https://doi.org/10.7275/f0gz-kc87

Hanna, G. S. (1975). Incremental reliability and validity of multiple-choice tests with an answer-until-correct procedure1. *Journal of Educational Measurement*, *12*(3), 175–178. https://doi.org/10.1111/j.1745-3984.1975.tb01019.x

Kane, M., & Moloney, J. (1978). The effect of guessing on item reliability under answer-until-correct scoring. *Applied Psychological Measurement*, *2*(1), 41–49. https://doi.org/10.1177/014662167800200104

Kastner, M., & Stangla, B. (2011). Multiple choice and constructed response tests: Do test format and scoring matter? *Procedia—Social and Behavioral Sciences*, *12*, 263–273. https://doi.org/10.1016/j.sbspro.2011.02.035

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*(2), 157–162. http://www.jstor.org/stable/1434513

Luce, R. D. (1959). *Individual choice behavior; a theoretical analysis*. Wiley, New York.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, *31*(3), 234–250. https://www.jstor.org/stable/1435268

Lyu, W., Bolt, D. M., & Westby, S. (2023). Exploring the effects of item-specific factors in sequential and IRTree models. *Psychometrika*. *88*, 745–775. https://doi.org/10.1007/s11336-023-09912-x

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement*, *27*(2), 133–144. https://www.jstor.org/stable/1434973

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(1), 1–97. https://doi.org/10.1007/BF03372160

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Slepkov, A. D., & Godfrey, A. T. K. (2019). Partial credit in answer-until-correct multiple-choice tests deployed in a classroom setting. *Applied Measurement in Education*, *32*(2), 138–150. https://doi.org/10.1080/08957347.2019.1577249

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577. https://doi.org/10.1007/BF02295596

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*(1), 39–55. https://doi.org/10.1111/j.2044-8317.1990.tb00925.x

Cite this article: Lu, Y., Fowler, J. and Cheng, Y. (2025). A Family of Sequential Item Response Models for Multiple-Choice, Multiple-Attempt Test Items. *Psychometrika*, 331–357. https://doi.org/10.1017/psy.2024.18