Article

Assessing Ability Recovery of the Sequential IRT Model With Unstructured Multiple-Attempt Data

Educational and Psychological Measurement
2022, Vol. 82(6) 1203–1224

DOI: 10.1177/00131644211058386
journals.sagepub.com/home/epm
SAGE

Ziying Li[1], A. Corinne Huggins-Manley[1], Walter L. Leite[1], M. David Miller[1] and Eric A. Wright[1]

Abstract

The unstructured multiple-attempt (MA) item response data in virtual learning environments (VLEs) are often from student-selected assessment data sets, which include missing data, single-attempt responses, multiple-attempt responses, and unknown growth ability across attempts, leading to a complex and complicated scenario for using this kind of data set as a whole in the practice of educational measurement. It is critical that methods be available for measuring ability from VLE data to improve VLE systems, monitor student progress in instructional settings, and conduct educational research. The purpose of this study is to explore the ability recovery of the multidimensional sequential 2-PL IRT model in unstructured MA data from VLEs. We conduct a simulation study to evaluate the effects of the magnitude of ability growth and the proportion of students who make two attempts, as well as the moderated effects of sample size, test length, and missingness, on the bias and root mean square error of ability estimates. Results show that the model poses promise for evaluating ability in unstructured VLE data, but that some data conditions can result in biased ability estimates.

Keywords

unstructured data, multiple attempts, virtual learning environment, sequential IRT models

[1]University of Florida, Gainesville, USA

Corresponding Author:
Ziying Li, University of Florida, 1602 Norman Hall, Gainesville, FL 32611, USA.
Email: ziying.li@ufl.edu

Virtual learning environments (VLEs) are online instructional systems in which learners and teachers interact for an educational purpose (Weller, 2007). Allowing or encouraging multiple attempts (MA) on assessment items is increasingly common in these VLEs (Abbakumov et al., 2018; Bergner et al., 2019). As opposed to some more traditional forms of MA assessment administration, such as an ''answer-until-correct'' item administration (Pressey, 1926), some VLEs allow students to attempt items multiple times at their own discretion while not requiring students, per say, to make multiple or even single attempts on items.

Due to these characteristics, MA data from VLEs can be considered unstructured as it often contains a mix of single- and multiple-attempt responses to items regardless of success on a first attempt, variability in the degree to which students' traits may increase across multiple attempts, and large amounts of missingness. In traditional MA data, there is often a fixed set of items in which each student makes at least one attempt and will make multiple attempts if the first attempt was incorrect, and there is often little time between attempts. However, in VLE MA data, students are often self-selecting to engage in assessments and may re-attempt an item at a much later date than the original attempt, if at all. Two characteristics of unstructured MA data from VLEs, namely the proportion of students making MA and the differential magnitudes of the changing ability across attempts per student, introduce variability in the data that has not been studied with respect to how to score the students. In addition, the unstructured MA data contain multiple types of missing data as some random missingness can come from items being randomly selected from the item bank to display to students who want to engage in the assessment and also some missing not at random (MNAR) data can be present when students choose to not respond to items presented or to only make one incorrect attempt. Hence, this study focuses on such item response data from these VLE-based, student-selected MA assessments, with a focus on being able to extract unbiased student ability estimates from such data, which can then be used for improving algorithms in the VLE (e.g., personalizing), informing instructors of student performance in the VLE, or conducting educational research from the VLE data.

Previous studies have applied classical test theory models (CTT), traditional item response theory models (IRT), and extended IRT models to evaluate data with MA to detect student learning (e.g., Abbakumov et al., 2018; Culpepper, 2014; Gönülates & Kortemeyer, 2015). For example, Attali and Powers (2010) followed the CTT framework to compare the observed scores between a three-attempt group and a single-attempt group on an MA test, finding that students' scores increased up to 19% in the group with three attempts. Abbakumov et al. (2018) extended the traditional IRT model by modeling the effect of attempts, finding that the growth of students' ability depended on the initial ability. Culpepper (2014) used a sequential Rasch model to estimate student ability in an up-to-five-attempt MA test, pointing out that student ability was constant at the first two attempts and then showed growth in following attempts. All these approaches analyzed their MA data either by adding an index to model the effect of attempts on student ability (e.g., Abbakumov et al.,

2018; Gönülates & Kortemeyer, 2015; Hung & Huang, 2019) or by separating the number of attempts into various groups (e.g., Attali & Powers, 2010; Cheng & Liu, 2016; Culpepper, 2014).

However, to our knowledge, neither these methods nor any other methods proposed in the literature for estimating ability from MA data have been evaluated in the context of unstructured MA item response data from VLEs. Since the data characteristics in previous research were not aligned with unstructured MA item response data from VLEs, it is critical to understand the performance of methods for estimating student ability from unstructured MA item response data. Otherwise, measurement models may result in biased or error-prone estimates. With the increasing popularity of online education and the increasing need for well-studied computational psychometrics methods, researchers and practitioners need appropriate measurement models to evaluate student ability from the unstructured VLE data.

One reason it is challenging to estimate student ability with unstructured MA data is because of the uncertainty about whether a particular student's ability is changing or staying constant across the repeated item attempts. It is easy and convenient for simulation studies to either assume constant ability or to assume growing ability within a particular degree across all simulees. But for applied studies, it is difficult for researchers to know the true ability change, leading to a challenge of model selection for appropriate ability estimates. In Culpepper's (2014) application study, he fit a unidimensional sequential Rasch model to a data set by classifying the number of attempts into different groups. After noticing student ability changed over attempts, he fit a multidimensional sequential Rasch model. Another example is from Bergner's et al. (2019) application study. Due to the unknown nature of true ability, they primarily analyzed model fit and the correlations between the estimated first-attempt scores and observed scores to evaluate the ability recovery of the sequential two-parameter logistic (2-PL) IRT model with three-attempt data, including MNAR data. As the true ability and ability growth are unknown in these applied studies, Culpepper's (2014) and Bergner's et al. (2019) results are bolstered by simulation work that supports the use of the approach in data similar to the applied data. However, as mentioned above, the unstructured MA data in VLEs are quite different than the data in previous MA studies, including those that used various versions of sequential measurement models, and hence simulations are needed to begin to understand how previously applied methods may work in such data environments.

The purpose of this study is to explore the ability of a sequential IRT model to recover unbiased ability estimates from unstructured MA data stemming from VLEs. We explore the model performance by manipulating four factors: the growth in student ability between attempts, the proportion of students making multiple attempts, the number of presented items, and the sample size. Also, we investigated the effect of MNAR data on the sequential IRT model performance. The outcomes of interest were bias and root mean square error (RMSE) surrounding ability estimates from the sequential IRT model. Hence, the research questions for this study are the following:

Research Question 1: Does the proportion of students with MA in items affect the bias and RMSE of ability estimates of the sequential IRT model?

Research Question 2: Does the growth in true ability affect the bias and RMSE of ability estimates of the sequential IRT model?

Research Question 3: Do the answers to research questions 1 and 2 depend on sample size or the number of presented items?

Research Question 4: Does the MNAR rate affect the bias and RMSE of ability estimates of the sequential IRT model?

The Sequential IRT Model and VLE Data

Sequential measurement models were proposed by Tutz (1990) and defined as

$$P\left(X_{ij}=r|\beta_i,\ \delta_{j1},\ \ldots,\ \delta_{jr}\right)=\prod_{s=1}^{r-1}\left(1-F\left(\delta_{js}-\beta_i\right)\right)F\left(\delta_{jr}-\beta_i\right), \tag{1}$$

where β_i is person i's ability, δ_{js} is item j's difficulty at the sth attempt, and r is the number of attempts. The function F can be defined in various ways and can come from one of the IRT or Rasch family models (Tutz, 1990). If we apply the binary Rasch model in a study, F is defined as

$$F\left(\delta_{js}-\beta_i\right)=\frac{\exp\left(\beta_i-\delta_{js}\right)}{1+\exp\left(\beta_i-\delta_{js}\right)}. \tag{2}$$

If we apply the 2-PL model (Birnbaum, 1968) into the sequential model framework, as Bergner et al. (2019) did, F is defined as

$$F\left(\delta_{js}-\beta_i\right)=\frac{\exp\left[\alpha_{js}\left(\beta_i-\delta_{js}\right)\right]}{1+\exp\left[\alpha_{js}\left(\beta_i-\delta_{js}\right)\right]}. \tag{3}$$

The logic of the sequential model's function is similar to the geometric random variable function, where $\prod_{s=1}^{r-1}(1-F(\delta_{js}-\beta_i))$ refers to the probability of the failure at the first s attempts and $F(\delta_{jr}-\beta_i)$ refers to the probability of the first success at the rth attempt. The details of distinguishing the differences between the sequential IRT models and the other IRT model extensions can be found in Bechger and Akkermans's (2001) research and Hemker et al.'s (2001) research, and they also mentioned the item parameterization in the sequential IRT models.

One reason to expect that sequential IRT models may perform well in unstructured MA data from VLEs is that these models have shown strong ability recovery performance in complex situations. Culpepper (2014) was the first application study using sequential models for a five-attempt test while defining the model as a sequential Rasch rating scale model. Also, he built the linear model for estimating student ability

at every attempt. Culpepper (2014) claimed that the sequential Rasch model can estimate student ability with less bias at every attempt. Bergner et al. (2019) conducted a simulation study for MA item response data and concluded that the sequential 2-PL IRT model was associated with low RMSE values in estimating ability even when the data contained missing values.

Given that sequential IRT models may appropriately estimate student ability under the MA response structure, we selected a sequential IRT model to evaluate whether this model can estimate student ability for the unstructured MA data. To estimate each student's various abilities during attempts in one model fitting, we selected the multidimensional 2-PL IRT model. The reason for selecting a 2-PL framework extension was that our previous work in the VLE platform established the appropriateness of a 2-PL framework underlying all of our item data (Xue et al., 2021). Therefore, the multidimensional sequential 2-PL IRT model is defined as

$$P\left(X_{ij}=r|\beta_{i1},\ \ldots,\beta_{ir},\ \delta_j\right)=\prod_{s=1}^{r-1}\left[1-F\left(\alpha_{js},\ \delta_{js},\ \beta_{i1},\ \beta_{i2}\right)\right]\left[F\left(\alpha_{jr},\ \delta_{jr},\ \beta_{i1},\ \beta_{i2}\right)\right], \tag{4}$$

where

$$F\left(\alpha_{js},\ \delta_{js},\ \beta_{i1},\ \beta_{i2}\right)=\frac{\exp\left[\alpha_{js}\left(\beta_{i1}-\delta_{js}\right)+\alpha_{js}\left(\beta_{i2}-\delta_{js}\right)\right]}{1+\exp\left[\alpha_{js}\left(\beta_{i1}-\delta_{js}\right)+\alpha_{js}\left(\beta_{i2}-\delta_{js}\right)\right]}. \tag{5}$$

Probability of Skipping Items and Skipping Rate in Our Prior Work With the VLE

In VLEs, the unstructured MA data contain MNAR data due to skipping items. Students studying in VLEs may skip some challenging practices or select the learning materials that they are interested in or familiar with, since they can self-determine their learning materials. In the VLE of this current study, there is a large bank of items associated with various domains of content. When a student decides to engage in assessment, 10 items are drawn randomly from the relevant item bank and provided to the student. On any given item, the student can make no attempt or one attempt. At a later date, the same student may decide to practice with assessment items again in the VLE, and they may receive an item they have attempted in the past, allowing for another attempt on that item. Hence, the data have a large mix of missingness, single attempt on items, and multiple attempts on items.

Our previous work found that student ability and item difficulty were systematically related to skipping items that were presented to students in the VLE (Xue et al., 2021). We used logistic regression to simulate the skipping item response data and showed that the percentage of missingness in the data generated for simulations was similar to the characteristics of the observational VLE data. Hence, for this study we defined the probability of skipping items in alignment with our previous work as

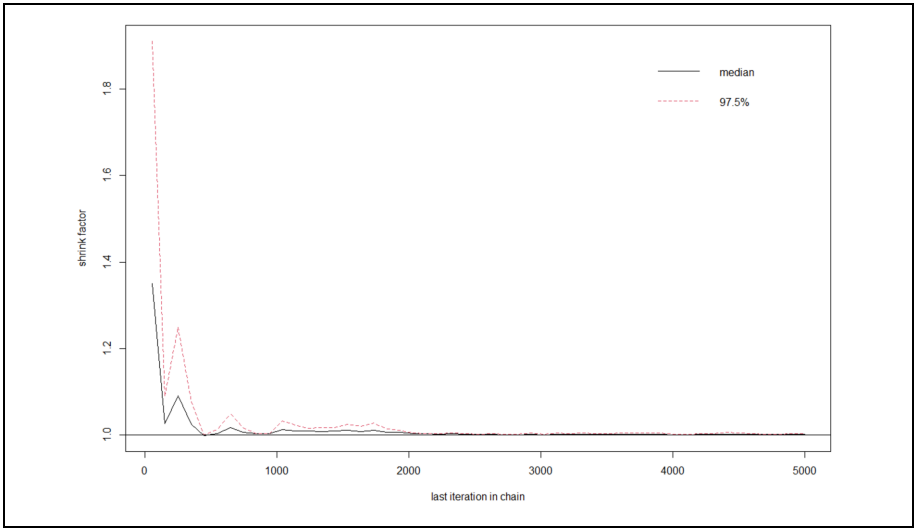

Figure 1. The Gelman-Rubin-Brooks Plot for Testing the Convergence Between and Within Chains in B_1 Sampling.

$$p_{ij} = P(skipping = 1) = \frac{\exp\left(B_1 Ability_i + B_2 Difficulty_j\right)}{1 + \exp\left(B_1 Ability_i + B_2 Difficulty_j\right)}, \tag{6}$$

where B_1 and B_2 are the coefficients for student ability and item difficulty. When $p_{ij}<0.5$, it means that Student i has a lower probability to skip Item j. In this case, we assumed Student i responded to Item j. On the contrary, if $p_{ij} \geq 0.5$, we assumed Student i skipped Item j.

In this study, we obtained B_1 and B_2 by using the random-walk-based Metropolis-Hastings algorithm with four chains in R 4.0 (R Development Core Team, 2020). The likelihood of skipping, Y_n, followed a Bernoulli distribution with probability where n is the sample size of 1,000. p_n was computed by Equation 6 where $Ability_n$ and $Difficulty_n$ were drawn from a standard normal distribution. The prior B_1 and B_2 were a normal distribution with mean of -0.10 and 0.12 and variance of 0.010 and 0.001, respectively, based on Xue et al. (2021). The proposal candidate $B_i^{'}$ was defined as

$$B_i^{'} = B_i + \epsilon, \tag{7}$$

where ϵ was a normal distribution with mean of 0 and variance of 0.10. Each chain conducted 10,000 iterations and discarded the first 5,000 burn-in values. Finally, the Gelman-Rubin-Brooks plots (Figures 1 and 2) of B_1 and B_2 sampling and each chain's Auto-Correlation-Function (ACF) plot (Figures 3 and 4) showed that our

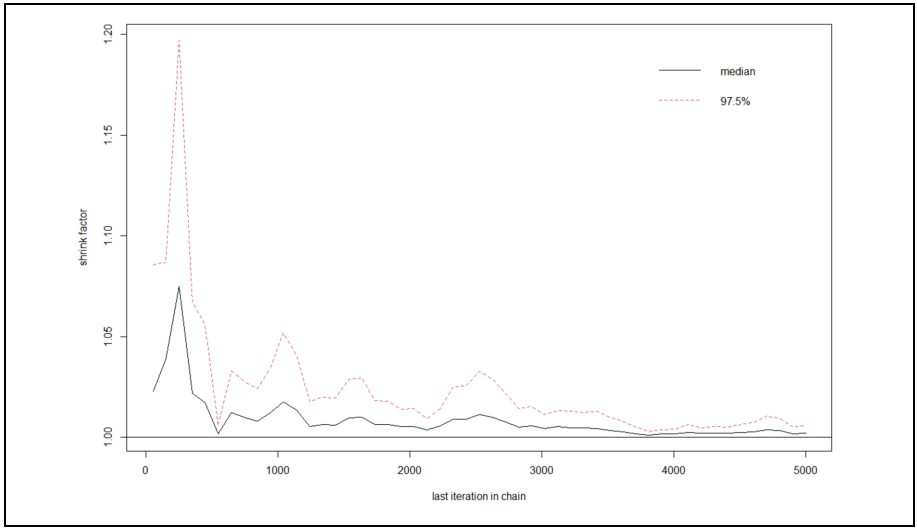

Figure 2. The Gelman-Rubin-Brooks Plot for Testing the Convergence Between and Within Chains in Sampling.

samples had appropriate posterior distributions. The Gelman-Rubin-Brooks plots were conducted by coda package (Plummer et al., 2020).

In addition, we considered the skipping rate when generating item response data. The skipping rate was defined as

$$R = \frac{The\ number\ of\ skipping\ items}{Total\ number\ of\ items}. \tag{8}$$

Suppose students should respond to all items, we can obtain the number of skipped items for each student by calculating every student's probability of skipping per item. A lower skipping rate R means that a student responds to more items. To be more specific, combining Equations 6 and 8, a higher ability student will respond to more items, leading to a higher probability of more re-attempting of items, or more MA. Our previous work has shown that higher ability students engage more with the VLE and show less item skipping and, thus, they have a higher probability of seeing items repeatedly in the system (Xue et al., 2021).

Current Simulation Study

A simulation study was conducted based on operational data from an online learning platform Math Nation (Lastinger Center for Learning, University of Florida, 2019), which is a statewide VLE that provides video resources about Algebra 1, Geometry, and Algebra 2 to assist upper middle school and high school students with learning

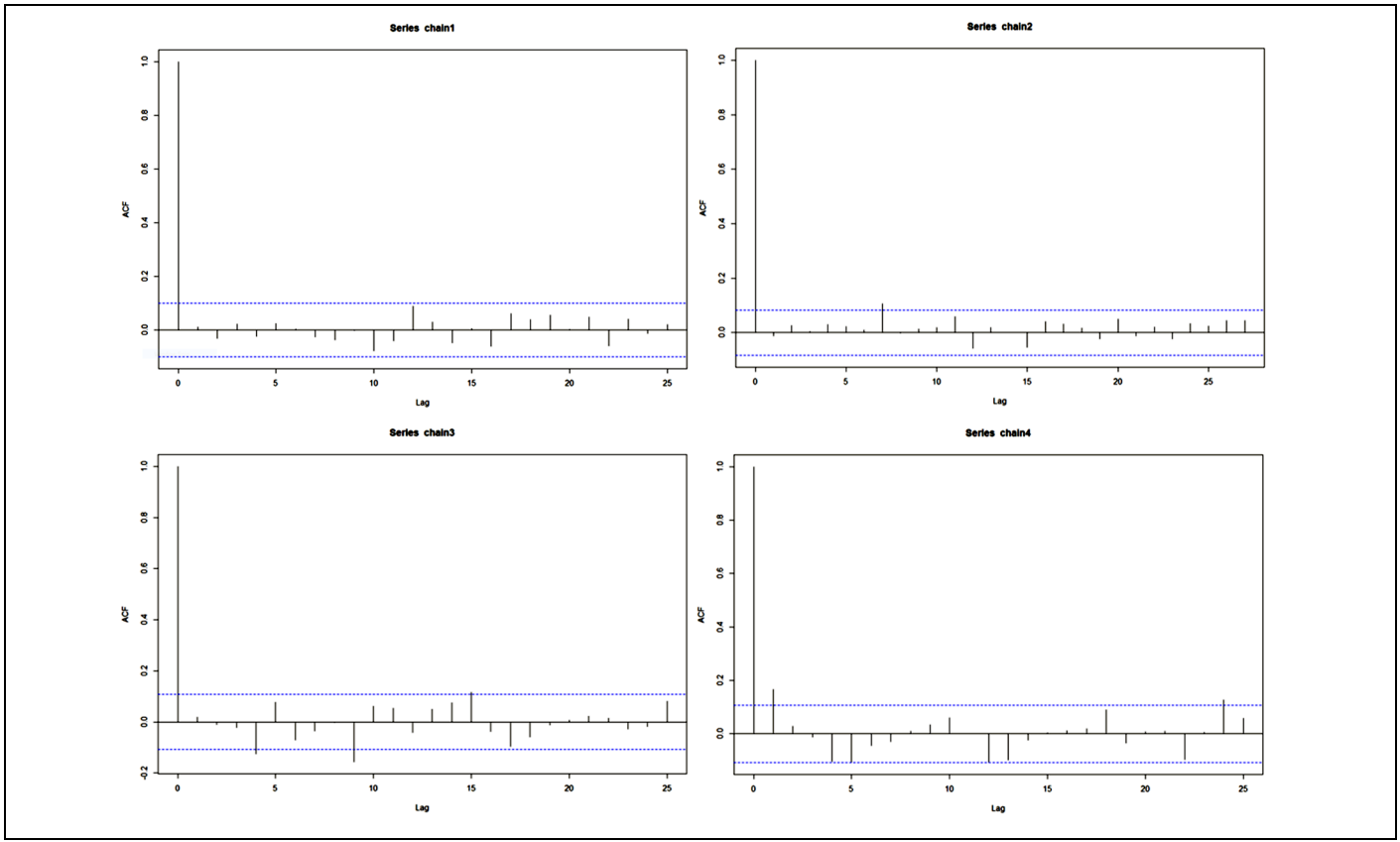

Figure 3. The Auto-Correlation-Function Plots of B_1 Sampling by Using the Metropolis-Hastings Algorithm.

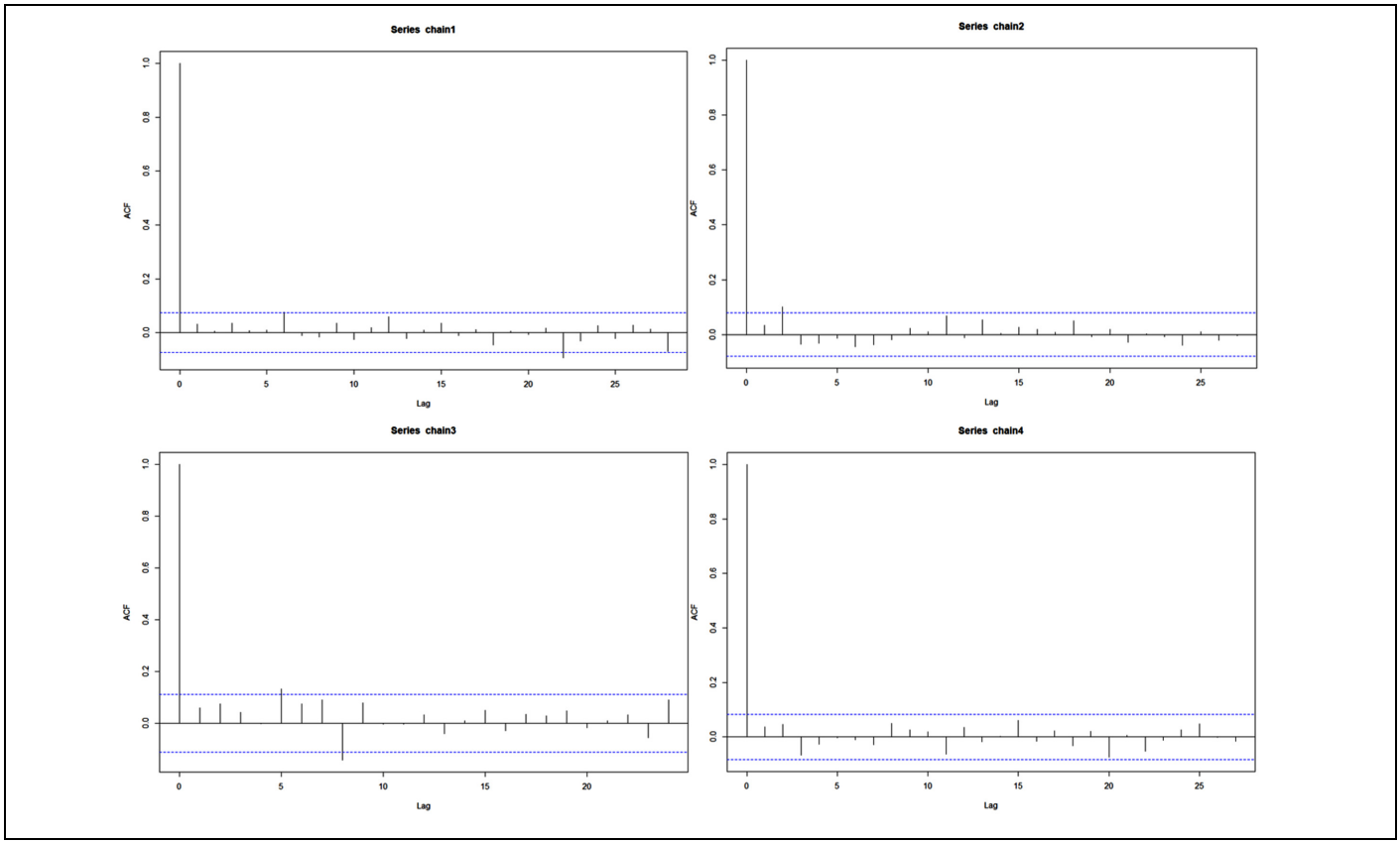

Figure 4. The Auto-Correlation-Function Plots of B_2 Sampling by Using the Metropolis-Hastings Algorithm.

in these mathematics subjects. This VLE allows students to determine their learning pace and select learning materials by themselves (Leite et al., 2019). Notably, the items presented to students neither relate to student ability nor the item selection in the next engagement with the VLE. Rather, the VLE randomly assigns items from their item bank after students select their learning materials. Teachers sometimes assign students to practice with the items, but they do not assign particular items as random ones are chosen by the system. Students can decide which items to engage with, and their responses do not affect the random selection of the presented items in the next engagement.

We obtained three operational data sets that were collected from Grade 8 to 10 students' VLE activity logs during the spring semester of 2018, the spring semester of 2020 before COVID-19 school closures, and the spring semester of 2020 after COVID-19 school closures. Each data set records the presented item numbers and students' scores at every attempt, which has some students re-attempting an item more than once. For those students who skip the presented items, their scores record as NAs in the data set. The simulation conditions in the current study mimic descriptive statistics from these three data sets, with the exception of the ability growth between attempts which is unknown in the operational data sets and assumed to vary across students as the time between attempts also varies. Due to the limited choices in each item, students can relatively easily guess the correct answer on the third or higher attempt of an item, assuming they recall their earlier attempts. Hence, this simulation study allowed students to attempt the same items no more than two times, as we only considered the first two attempts in our operational data to be data that can help to evaluate student ability.

Simulation Factors

The proportion of students making two attempts on items was created as five levels, which were 1, 0.8, 0.6, 0.4, and 0.2. In the three operational data sets, the proportions of students making MA on at least one item varied were 13%, 24%, and 39%. In other words, the proportion of students making MA on items depends on the sample of data. Therefore, we set five levels with equal intervals to explore the effect of proportion of students making MA on the ability parameter recovery.

We set three levels and an additional baseline level for the simulation factor of ability growth between attempts, σ_λ, which were (a) (0, 0) as a baseline level, that is, student ability increases 0 standard deviations (*SD*) from the first to second attempt; (b) (0, 0.5 *SD*), that is, students who have a less than 0.5 skipping rate (Equation 8) increase 0.5 *SD* ability on their second attempt while abilities of those with higher skipping rates does not change; (c) (0, 1 *SD*), that is, students who have a less than 0.5 skipping rate increase 1 *SD* ability on their second attempt while others' do not show ability change; (d) (0.5 *SD*, 1 *SD*), that is, students who have a less than 0.5 skipping rate increase 1 *SD* ability while others' abilities increase 0.5 *SD* on their second attempt. As we discussed the skipping rate in an above section, our previous work (Xue et al., 2021) found that higher ability students had lower skipping rates

on average, and thus may attempt more items and have more multiple-attempt items. Hence, students with a lower skipping rate had a higher magnitude in ability growth than those who had a higher skipping rate. We selected 0.5 *SD* units and 1 *SD* units from previous MA studies (e.g., Feinberg et al., 2015) to represent a moderate and strong magnitude in ability growth across two attempts, respectively.

The sample size was set as three levels, which were 500, 1,000, and 5,000. The sample sizes of three operational data sets were all larger than 3,000. It is a common situation in statewide or nationwide VLEs that the sample size is larger than 3,000. Thus, we set a level of 5,000 to represent those VLE data sets' sample sizes. The level of 1,000 is a common level of sample size in many traditional IRT studies, and thus some researchers also selected this level of sample size to conduct MA studies with sequential IRT models (e.g., Bergner et al., 2019). Although research about the level of the sample size of 1,000 has been evaluated in traditional IRT models, the complicated characteristics of unstructured MA data in VLEs include a small number of item responses per student compared with the total number of items, various proportions of students making MA on items, and a nonignorable proportion of missing data. We decided to set 1,000 as one of the simulation sample size levels in our study for providing evidence to support the sample size selection in future MA application research in smaller VLE data sets. The level of 500 is also a common condition in other online educational learning studies (e.g., Hamutoglu et al., 2020; Jones et al., 2021). Hence, we set a level of 500 to represent small sample sizes in VLEs.

We created three levels for the simulation factor of the number of presented items, which were 5, 15, and 30. The VLE under study presents 10 items at a time, and most students engage with approximately one to three sets of items, with some skipping present. Thus, we set a level of 30 items as a baseline, with 15 items as a moderate amount of item engagement, and 5 items for students who are less engaged.

In total, there were 180 conditions, each of which were evaluated under 100 iterations in this study.

Data Generation

In this section, we mimicked the characteristics of the unstructured MA data in VLEs based on the operational data sets. We randomly sampled items for every student and randomly selected a proportion of students to make a second attempt on items. The response data for each attempt was generated separately with the 2-PL IRT model (Equation 3). The initial latent ability β_i was sampled from a normal distribution with a mean of 0 and a standard deviation of 1. The second-attempt ability $\beta_i^{'}$ was computed from $(\beta_i + \lambda)$, where λ represents ability change, and was sampled from a normal distribution with a mean of 0 and a standard deviation of σ_λ. Below we provide an example of the data generated under the condition that σ_λ is (0, 0.5 *SD*), the sample size is 1,000, the number of presented items is 15, and the proportion of students making MA on items is 0.2.

At the first step, we randomly selected 15 items for every student who might skip items based on Equation 6. We re-sampled randomly from the remaining unselected

items until students had at least five responses for mimicking the characteristics of the operational data. In a special case of students who skipped all items, we randomly selected two items to record their responses to keep those students in the database. Without this catch, many students went completely missing on the five-item test, making for incomparable simulation results across conditions. Next, we randomly selected 200 students out of 1,000 to make a second attempt on items. The proportion of re-attempt items depended on those selected students' skipping rates (see Equation 8). Suppose a student responded to five items at the first attempt and their skipping rate was 90%, then we randomly selected one item (5×10%) as the re-attempt item from the first-attempt items. To keep the number of presented items consistent across conditions, we re-sampled 14 items (15 − 1) from the un-selected items. At the third step, the selected students who had a less than 0.5 skipping rate (Equation 8) increased 0.5 *SD* in their ability at the second attempt while the remaining selected students had 0 ability growth.

For the generation of item parameters, item discriminations, α_{js}, were sampled from a uniform distribution with a range from 0.8 to 2.19 based on the range of the operational data sets. The item difficulties, δ_{js}, were generated from a standard normal distribution, which was a close approximation to the operational item difficulties. Notably, item parameters were known and set to be constant in the generation of data across the two attempts. These unbiased item parameters were obtained from a previous study (Xue et al., 2021), which used neural networks to remove bias from MNAR data in item parameters estimates from the 2-PL IRT model using the same source data as us. The data were generated in R 4.0 (R Development Core Team, 2020).

Data Analysis and Evaluation Criteria

We fit the multidimensional sequential 2-PL IRT model (Equations 4 and 5) while, notably, constraining item parameters to be equal across attempts. The models were fit in the MIRT package (Chalmers et al., 2020) in R.

To evaluate the recovery of latent ability β_i and β_i', the bias and the RMSE were calculated. The bias is used to assess the difference between the estimated ability and the true ability and is defined as

$$bias = \frac{\sum_{i=1}^{n}\left(\widehat{\beta}_i - \beta_i\right)}{n} \tag{9}$$

where n is the number of repeated times, and $\hat{\beta}_i$ is the estimated latent ability. A smaller magnitude of bias represents a better recovery of ability by the sequential model. An unbiased estimate will be equal to 0 after averaging over the iterations (Feinberg & Rubright, 2016).

RMSE is used to compare the deviations of the estimated abilities, and is defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(\widehat{\beta}_i - \beta_i\right)^2}{n-1}}. \quad (10)$$

To investigate the research questions, four factorial analysis of variance (ANOVA) models were used to analyze the main and interaction effects of the four simulation factors and the MNAR rate variable on the bias and RMSE, treated separately. The MNAR rate is defined as

$$MNAR = \frac{the\ number\ of\ skipping\ items\ in\ each\ iteration\ dataset}{the\ total\ number\ of\ presented\ items\ in\ each\ iteration\ dataset}. \quad (11)$$

Results

For all 180 conditions, the bias of the first-attempt ability estimates ranged from -0.77 to -0.05, and the RMSE of initial ability estimates ranged from 0.07 to 0.79. For the second-attempt ability estimates, the bias ranged from -0.82 to 0.44, and its RMSE ranged from 0.04 to 0.86. These results indicate that the multidimensional sequential 2-PL IRT model has a good recovery in student ability under some conditions but not all conditions. Prior to submitting these results to ANOVA models, we checked and upheld the normality, independence, and linearity assumptions.

As shown in Tables 1 and 2, the proportion of students making MA on items had a statistically significant main effect on bias and RMSE of the first-attempt ability estimates ($F_{T1_{bias},prop}(4) = 1943.792, p < 2\times10^{-16}$, $\eta^2 = 0.50$; $F_{T1_{RMSE},prop}(4) = 1237.393$, $p < 2\times10^{-16}$, $\eta^2 = 0.40$). As the proportion increased, the absolute value of bias and RMSE in the first-attempt ability estimation increased (see Figures 5 and 6). When the proportion of students making MA on items was 0.2, the bias and RMSE of the first-attempt ability, to some extent, was smaller than in cases of larger proportions of MA.

Tables 3 and 4 showed that for the second-attempt ability estimates there was also a main effect on the bias and RMSE of the proportion of students making MA on items ($F_{T2_{bias},prop}(4) = 4316.672, p < 2\times10^{-16}$, $\eta^2 = 0.18$; $F_{T2_{RMSE},Prop}(4) = 92.244$, $p < 2\times10^{-16}$, $\eta^2 = 0.11$). Notably, the higher the proportion of students making MA on items the larger the bias of the initial ability estimates but the lower the bias of the second-attempt ability estimates (Figure 5), which makes sense as more MA is needed to have sufficient data to have unbiased estimates for the second-attempt ability. There was a similar finding for RMSE, as shown in Figure 6. When 20% of students made MA on items, we found some of the smallest bias and RMSE values for the initial ability estimation but also some of the largest bias and RMSE values for the second-attempt ability estimation. Conversely, when the proportion of students making MA on items was no less than 0.8, the bias and RMSE of the ability estimates were relatively large for the initial-attempt ability but relatively small for the second-attempt ability (see Figures 5 and 6). Therefore, having a moderate proportion of students making MA on items, such as 0.4 and 0.6, resulted in the most

Table 1. ANOVA for Bias of the Initial Ability Estimate.

Predictors	Sum of Squares	*df*	*M* square	*F*	*p*	Eta square
Proportion	2.240	4	0.560	1,943.792	<2e-16***	0.50
Growth	0.184	3	0.061	212.369	<2e-16***	0.04
Sample size	0.301	2	0.150	521.633	<2e-16***	0.07
Items	1.143	2	0.572	1,983.581	<2e-16***	0.25
Missing not at random rate	0.045	1	0.045	157.568	<2e-16***	0.01
Sample × Items	0.420	4	0.105	364.670	<2e-16***	0.09
Sample × Growth	0.001	6	0.000	0.488	0.816	1.87e-04
Items × Growth	0.012	6	0.002	7.162	2.55e-06***	2.75e-03
Proportion × Growth	0.023	12	0.002	6.554	2.14e-08***	5.03e-03
Proportion × Sample	0.016	8	0.002	6.986	3.26e-07***	3.57e-03
Proportion × Items	0.061	8	0.008	26.323	<2e-16***	0.01
Grow × Sample × Items	0.008	12	0.001	2.304	0.013*	1.77e-03
Proportion × Sample × Items	0.023	16	0.001	4.991	2.45e-07***	5.11e-03
Residuals	0.027	95	0.000			

Note. The presented model was determined by removing the interaction terms that had neither statistically significant effects on bias of the initial ability estimates nor statistically significant effects on the bias of the second-attempt ability estimates.
*$p<0.05$. ***$p<0.001$.

Table 2. ANOVA for Root Mean Square Error of the Initial Ability Estimate.

Predictors	Sum of Squares	*df*	*M* square	*F*	*p*	Eta square
Proportion	1.843	4	0.461	1,237.393	<2e-16***	0.40
Growth	0.185	3	0.062	165.827	<2e-16***	0.04
Sample size	0.471	2	0.235	631.997	<2e-16***	0.10
Items	1.260	2	0.630	1,692.545	<2e-16***	0.27
Missing not at random rate	0.037	1	0.037	98.766	2.24e-16***	8.01e-03
Sample × Items	0.563	4	0.141	378.084	<2e-16***	0.12
Sample × Growth	0.001	6	0.000	0.440	0.850	2.14e-04
Items × Growth	0.016	6	0.003	6.978	3.62e-06***	3.39e-03
Proportion × Growth	0.021	12	0.002	4.730	5.20e-06***	4.60e-03
Proportion × Sample	0.033	8	0.004	11.025	6.84e-11***	7.15e-03
Proportion × Items	0.066	8	0.008	22.265	<2e-16***	0.01
Grow × Sample × Items	0.010	12	0.001	2.139	0.021*	2.08e-03
Proportion × Sample × Items	0.051	16	0.003	8.585	1.88e-12***	0.01
Residuals	0.035	95	0.000			

Note. *$p<0.05$. ***$p<0.001$.

acceptable overall amount of bias and RMSE when simultaneously considering the initial and second-attempt ability estimation.

Regardless of the initial or the second-attempt ability estimates, Tables 1 to 4 showed that the impact of the proportion of students making MA neither depended

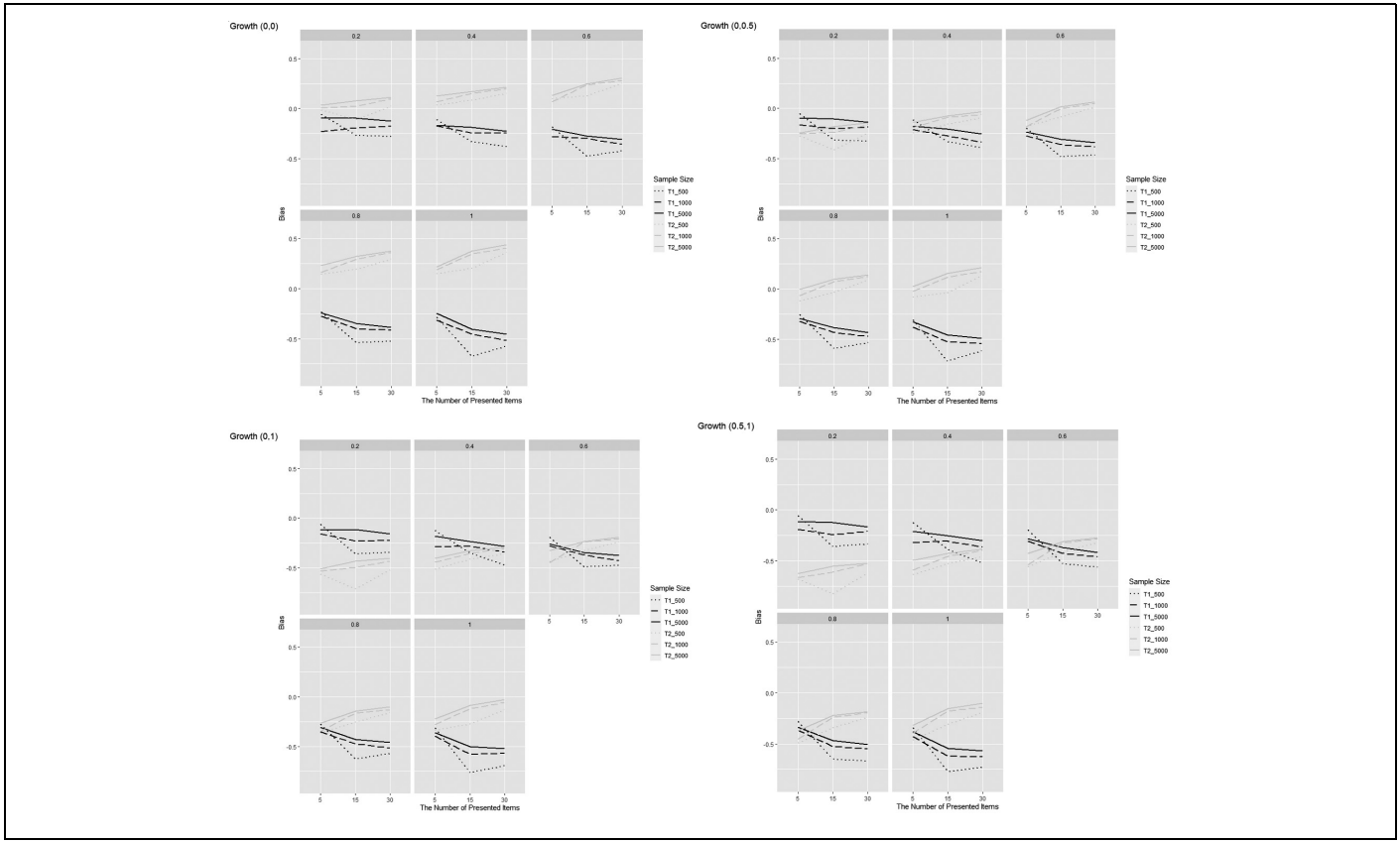

Figure 5. The Panel of Three-Way Comparison of the First- and Second-Attempt Ability Estimate Bias, When Controlling the Growth Ability.
Note. Each graph in this panel represents the relationship of percentage of multiple-attempt students with sample sizes and the number of presented items. For example, 0.2 means that 20% of students made two attempts on items. Each graph illustrates the change of bias as the sample size and the number of items increased among various proportions of students making two attempts on items.

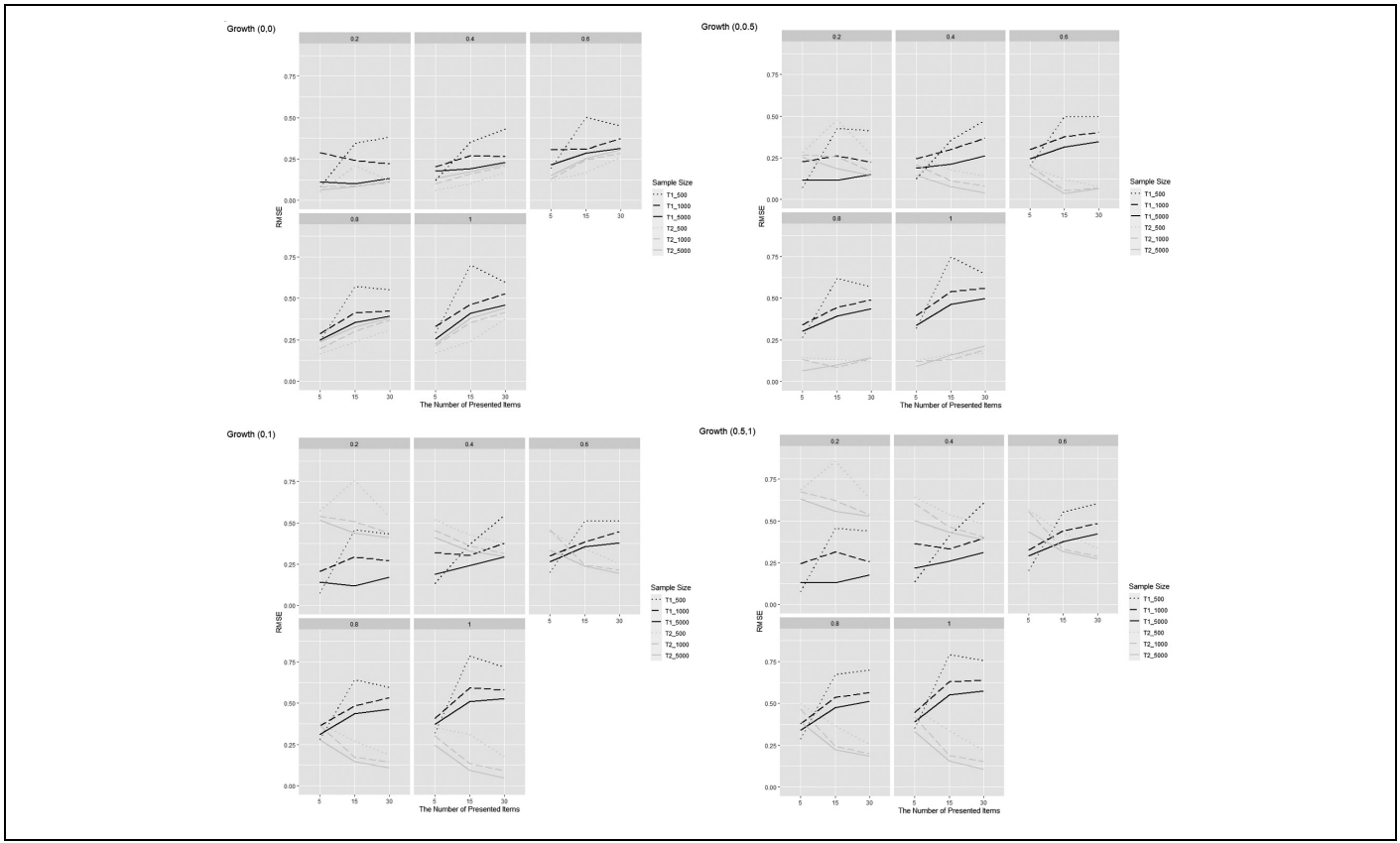

Figure 6. The Panel of Three-Way Comparison of the Change of Root Mean Square Error of the First- and Second-Attempt Ability Estimates When Controlling the Growth Ability.

Note. Each graph in this panel represents the change of root mean square error as the sample size and the number of items increased among various proportions of students making two attempts on items.

on the sample size nor the number of presented items (see negligible Eta-squared effect sizes).

The magnitude of ability growth had small and even close-to-zero main effects on the bias and RMSE of the first-attempt ability estimates. However, it had statistically and practically significant main effects on the bias and RMSE of the second-attempt ability estimates ($F_{T2_{bias},\,Growth}(3) = 22905.336, p < 2\times10^{-16},\ \eta^2 = 0.71$; $F_{T2_{RMSE},\,Growth}(3) = 435.106,\ p < 2\times10^{-16},\ \eta^2 = 0.38$). The main finding here was that when student ability did not grow across the two attempts, the second-attempt ability cannot be well estimated, presumably because it was not distinguished from the initial-attempt ability. The gray lines in Figure 5 showed that when some portion of students had moderate growth in their second-attempt ability and the rest of the students' abilities did not change across attempts (Growth [0,0.5]), the bias was smaller and close to 0 as compared with other ability growth conditions. Similar descriptions can be made of the second-attempt ability estimate results and their associated RMSE outcomes (see gray lines in Figure 6).

In addition, the magnitude of ability growth had an interaction effect with the number of presented items on the RMSE of the second ability estimates ($F_{T2_{RMSE},\ Growth^{*}\,Items}(6) = 56.166,\ p < 2\times10^{-16},\ \eta^2 = 0.10$). The gray lines in Figure 6 described that the RMSE of the second-attempt ability was large in a short test (e.g., 5 items) when the growth ability was (0, 1) or (0.5, 1). Such results indicated that a better estimation of the second-attempt ability needs a test with no less than 15 presented items, particularly when some examinees may have large growth in ability across attempts.

The MNAR rate ranged from 0.24 to 0.43. Tables 1 to 4 showed that the MNAR rate had close-to-zero effects on the bias and RMSE regardless of the initial or the second-attempt ability estimates.

As a final finding of practical and statistical significance, there was an interaction between sample size and test length in their effects on the bias and RMSE of ability estimates, shown in Tables 1 and 2 ($F_{T1_{bias},\ Sample^{*}\,Items}(4) = 364.670,\ p < 2\times10^{-16},\ \eta^2 = 0.09$; $F_{T1_{RMSE},\ Sample^{*}\,Items}(4) = 378.084,\ p < 2\times10^{-16},\ \eta^2 = 0.12$). Also, each of these simulation factors has a main effect on the bias and RMSE of the initial ability estimates ($F_{T1_{bias},\ Sample}(2) = 521.633,\ p < 2\times10^{-16},\ \eta^2 = 0.07$; $F_{T1_{bias},\ Items}(2) = 1983.581,\ p < 2\times10^{-16},\ \eta^2 = 0.25$; $F_{T1_{RMSE},\ Sample}(2) = 631.997,\ p < 2\times10^{-16},\ \eta^2 = 0.10$; $F_{T1_{RMSE},\ Items}(2) = 1692.545,\ p < 2\times10^{-16},\ \eta^2 = 0.27$). As shown in Figures 5 and 6, under conditions with sample size no less than 1,000 and number of presented items no less than 15, we found relatively less bias and RMSE results. For the second-attempt ability estimates, Tables 3 and 4 showed that the number of presented items had a main effect on the estimated bias ($F_{T2_{bias},\ Items}(2) = 2922.553,\ p < 2\times10^{-16},\ \eta^2 = 0.06$), where the bias mostly trended toward zero as the number of presented items increased.

Discussion

The purpose of this study was to evaluate the ability parameter recovery of the multidimensional sequential 2-PL IRT model when applied to unstructured MA item

Table 3. ANOVA for Bias of the Second Ability Estimate.

Predictors	Sum of Squares	*df*	*M* square	*F*	*p*	Eta square
Proportion	2.425	4	0.606	4,316.672	<2e-16***	0.18
Growth	9.652	3	3.217	22,905.336	<2e-16***	0.71
Sample size	0.334	2	0.167	1,190.632	<2e-16***	0.02
Items	0.821	2	0.411	2,922.553	<2e-16***	0.06
Missing not at random rate	0.040	1	0.040	282.324	<2e-16***	2.93e-03
Sample × Items	0.052	4	0.013	93.015	<2e-16***	3.87e-03
Sample × Growth	0.002	6	0.000	2.439	0.031*	1.52e-04
Items × Growth	0.013	6	0.002	14.848	6.63e-12***	9.26e-04
Proportion × Growth	0.036	12	0.003	21.163	<2e-16***	2.64e-03
Proportion × Sample	0.017	8	0.002	15.197	3.75e-14***	1.26e-03
Proportion × Items	0.077	8	0.010	68.189	<2e-16***	5.67e-03
Grow × Sample × Items	0.002	12	0.000	1.382	0.188	1.72e-04
Proportion × Sample × Items	0.032	16	0.002	14.334	<2e-16***	2.38e-03
Residuals	0.013	95	0.000			

Note. *$p<0.05$. ***$p<0.001$.

Table 4. ANOVA for Root Mean Square Error of the Second Ability Estimate.

Predictors	Sum of Squares	*df*	*M* square	*F*	*p*	Eta square
Proportion	0.541	4	0.135	92.244	<2e-16***	0.11
Growth	1.914	3	0.638	435.106	<2e-16***	0.38
Sample size	0.130	2	0.065	44.364	2.47e-14***	0.03
Items	0.128	2	0.064	43.766	3.37e-14***	0.03
Missing not at random rate	0.000	1	0.000	0.229	0.634	6.72e-05
Sample × Items	0.033	4	0.008	5.549	0.000***	6.53e-03
Sample × Growth	0.151	6	0.025	17.151	2.44e-13***	0.03
Items × Growth	0.494	6	0.082	56.166	<2e-16***	0.10
Proportion × Growth	1.293	12	0.108	73.469	<2e-16***	0.26
Proportion × Sample	0.038	8	0.005	3.198	0.003**	7.53e-03
Proportion × Items	0.061	8	0.008	5.231	1.97e-05***	0.01
Grow × Sample × Items	0.015	12	0.001	0.832	0.617	2.94e-03
Proportion × Sample × Items	0.048	16	0.003	2.056	0.017*	9.68e-03
Residuals	0.139	95	0.002			

Note. *$p<0.05$. **$p<0.01$. ***$p<0.001$.

response data in VLEs. Results from the simulation study demonstrated that the multidimensional sequential 2-PL IRT model can recover ability parameters in some data contexts that contain mixed single and multiple attempts, various ability growth, and MNAR data. Importantly, even when the MNAR rate in the unstructured MA

data reaches 42.9%, the multidimensional sequential 2-PL IRT model can recover ability parameters in a similar manner to when lower MNAR rate is present. However, some data conditions do result in increased bias and RMSE compared with other conditions.

We found that the recovery of ability parameters from the multidimensional 2-PL IRT model is affected by the proportion of students making MA on items, sample size, and the number of presented items. This is a positive finding because those factors are relatively manipulable in practice. When an unstructured MA data set contains a moderate proportion of students making multiple attempts on items (i.e., 40%–60%), the multidimensional sequential 2-PL IRT model has a strong recovery of the first-attempt and second-attempt ability parameters. Generally, if the unstructured MA data set has more than 60% of students making multiple attempts on items, the model can estimate the second-attempt ability with less bias to explore student ability growth with a known prior ability.

Based on our results, we recommend applying the multidimensional sequential 2-PL IRT model with a sample size of more than 1,000 and a test length of no less than 15 items to obtain acceptable estimates for both the first- and second-attempt ability. This study demonstrated that sample size and the number of presented items play nonignorable roles in obtaining unbiased estimates of ability. These findings related to the effects of sample sizes and test length on parameter recovery are redundant with traditional IRT research, but serve as a reminder that these features matter in both traditional measurement data as well as complex MA data from technology-enhanced platforms. For researchers and practitioners considering using a multidimensional sequential IRT model to estimate ability from the unstructured MA data in VLEs, we recommend paying close attention to sample sizes and the number of presented items.

Notably, we found that the magnitude of growth ability and the MNAR rate, which are not under the control of researchers and practitioners, have limited impacts, and in some cases no impact, on the recovery of ability parameters from the multidimensional sequential 2-PL IRT model. The only impact, at least based on our study outcomes, is that the model has a better performance in second-attempt ability recovery when the unstructured MA data set contains students with moderate growth ability and the rest without growth. But such an effect can be adjusted, to some extent, by setting the number of presented items to no less than 15. We also recommend setting a maximum or minimum amount of time between attempts to include in the data, as likely this can reduce variability in the magnitude of ability growth between attempts.

This study shows some promise for researchers and practitioners to consider using the multidimensional sequential 2-PL IRT model in their VLE data containing some MA to obtain unbiased ability estimates, assuming of course that the data are similar to our simulated conditions in which we found acceptable levels of bias and RMSE. In cases where the data conditions of a VLE are quite different from our studied conditions or align with our conditions in which bias and RMSE were larger, we encourage conducting a simulation like ours but with adjusted settings and comparing the

model performance to other options for ability estimation. Indeed, we found our simulation very helpful in understanding how we can estimate ability in our data, and we believe other practitioners and researchers of VLE data may find the same benefits with adjusted simulations.

This discussion of data generation in our simulation study brings up recommendations for future studies. For every data set we generated in our simulation, we grouped students into two types, those who make a single attempt on items and those who make two attempts on items. Meanwhile, we related the number of multiple-attempt items to student ability and item difficulty. In other words, we simulated the amount of MA as a property of the interaction of students with particular items. This is supported by our operational data, which shows that students' MA on items relates to their ability and item difficulty. However, others may want to consider simulating an environment in which MA is treated as a property of item-based features only (i.e., group items into those that have MA and those that do not) if that better represents their VLE unstructured MA item response data. Therefore, a recommendation for future studies is to evaluate the recovery ability of the multidimensional sequential 2-PL IRT model by simulating MA data based on item-level differences rather than person-item level interactions. Another recommendation for future studies is to consider the effect of more than two attempts on the model's recovery of ability. Based on the characteristics of the operational data, we allowed students to attempt items up to twice. However, others may want to consider the ability recovery of the multidimensional sequential 2-PL IRT model with more than two attempts, such as in an ''answer-until-correct'' format. Such studies can enhance the target precision of the sequential IRT model application in broad VLEs with unstructured MA item response data.

Ultimately, our study results indicate acceptable recovery of ability estimates under some conditions, but not all conditions, from the multidimensional sequential 2-PL IRT model fit to unstructured MA data in our simulated VLE-like environments. Although the proportion of students making multiple attempts on items, the sample size, and the number of presented items have effects on the ability parameter recovery of the multidimensional sequential 2-PL IRT model, these factors are relatively manipulable in practice. For those factors that are not under the control of measurement practitioners and researchers (i.e., the magnitude of growth ability and the MNAR rate), they mostly do not affect the ability estimation from the multidimensional sequential 2-PL IRT model with unstructured MA data. Only the magnitude of growth ability has an effect on the second-attempt ability estimates, but it can be reduced by adjusting test length. Therefore, we conclude that the multidimensional sequential 2-PL IRT model poses much promise for evaluating student ability and change in ability in VLE systems.

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C160004 to the University of Florida. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

ORCID iDs

Ziying Li https://orcid.org/0000-0003-4375-8332
Walter L. Leite https://orcid.org/0000-0001-7655-5668
M. David Miller https://orcid.org/0000-0002-3506-1167

References

Abbakumov, D., Desmet, P., & Van Den Noortgate, W. (2018). Measuring student's proficiency in MOOCs: Multiple attempts extensions for the Rasch model. *Heliyon*, *4*(12), 1–15. https://doaj.org/article/67aae6b2353e4701a78b117c31b821bf

Attali, Y., & Powers, D. (2010). Immediate feedback and opportunity to revise answers to open-ended questions. *Educational and Psychological Measurement*, *70*(1), 22–35. http://doi.org/10.1177/0013164409332231

Bechger, T. M., & Akkermans, W. (2001). A note on the equivalence of the graded response model and the sequential model. *Psychometrika*, *66*(3), 461–464. https://link.springer.com/article/10.1007/BF02294445

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

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In F.M. Lord & M.R. Novick (Eds.). *Statistical Theories of Mental Test Scores* (pp. 397–479). Addison-Wesley.

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K., Falk, C. F., Meade, A., Schneider, L., King, D., Liu, C. W., & Oguzhan, O. (2020, October 31). Mirt: Multidimensional item response theory. *GitHub*. https://CRAN.R-project.org/package=mirt

Cheng, Y., & Liu, C. (2016). A short note on the relationship between pass rate and multiple attempts. *Journal of Educational Measurement*, *53*(4), 431–447. https://doi.org/10.1111/jedm.12124

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

Feinberg, R. A., Raymond, M. R., & Haist, S. A. (2015). Repeat testing effects on credentialing exams: Are repeaters misinformed or uninformed? *Educational Measurement: Issues and Practice*, *34*(1), 34–39. https://doi.org/10.1111/emip.12059

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36–49. https://doi.org/10.1111/emip.12111

Gönülates, E., & Kortemeyer, G. (2015). A new item response theory model for open-ended online homework with multiple allowed attempts. *Physical Review Special Topics-Physics Education Research*, *3*, Article 06287v1. http://arxiv.org/abs/1503.06287

Hamutoglu, N. B., Gemikonakli, O., Duman, I., Kirksekiz, A., & Kiyici, M. (2020). Evaluating students experiences using a virtual learning environment: Satisfaction and preferences. *Educational Technology Research and Development*, *68*(1), 437–462. https://doi.org/10.1007/s11423-019-09705-z

Hemker, B. T., Andries van der Ark, L., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika*, *66*(4), 487–506. https://doi.org/10.1007/BF02296191

Hung, S., & Huang, H. (2019). A sequential process model for cognitive diagnostic assessment with repeated attempts. *Applied Psychological Measurement*, *43*(7), 495–511. https://doi/10.1177/0146621618813111

Jones, D., Lotz, N., & Holden, G. (2021). A longitudinal study of virtual design studio (VDS) use in STEM distance design education. *International Journal of Technology and Design Education*, *31*, 839–865. https://doi.org/10.1007/s10798-020-09576-z

Lastinger Center for Learning, University of Florida. (2019). *Algebra Nation*. http://lastingercenter.com/portfolio/algebra-nation-2/

Leite, W. L., Cetin-Berber, D. D., Huggins-Manley, A. C., Collier, Z. K., & Beal, C. R. (2019). The relationship between Algebra Nation usage and high-stakes test performance for struggling students. *Journal of Computer Assisted Learning*, *35*(5), 569–581. https://doi.org/10.1111/jcal.12360

Plummer, M., Best, N., Cowles, K., Vines, K., Sarkar, D., Bates, D., Almond, R., & Magnusson, A. (2020, September 30). *Package ''coda.''* https://CRAN.R-project.org/package=coda

Pressey, S. L. (1926). A simple apparatus which gives tests and scores and teaches. *School and Society*, *23*(586), 373–376.

R Development Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

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

Weller, M. (2007). *Virtual learning environments: Using, choosing and developing your VLE*. Routledge.

Xue, K., Huggins-Manley, A. C., & Leite, W. L. (2021). Semisupervised learning method to adjust biased item difficulty estimates caused by nonignorable missingness in a virtual learning environment. *Educational and Psychological Measurement*.