Article

Educational and Psychological Measurement
2022, Vol. 82(3) 539–567

DOI: 10.1177/00131644211020494
journals.sagepub.com/home/epm
SAGE

Semisupervised Learning Method to Adjust Biased Item Difficulty Estimates Caused by Nonignorable Missingness in a Virtual Learning Environment

Kang Xue[1], Anne Corinne Huggins-Manley[2] and Walter Leite[2]

Abstract

In data collected from virtual learning environments (VLEs), item response theory (IRT) models can be used to guide the ongoing measurement of student ability. However, such applications of IRT rely on unbiased item parameter estimates associated with test items in the VLE. Without formal piloting of the items, one can expect a large amount of nonignorable missing data in the VLE log file data, and this is expected to negatively affect IRT item parameter estimation accuracy, which then negatively affects any future ability estimates utilized in the VLE. In the psychometric literature, methods for handling missing data have been studied mostly around conditions in which the data and the amount of missing data are not as large as those that come from VLEs. In this article, we introduce a semisupervised learning method to deal with a large proportion of missingness contained in VLE data from which one needs to obtain unbiased item parameter estimates. First, we explored the factors relating to the missing data. Then we implemented a semisupervised learning method under the two-parameter logistic IRT model to estimate the latent abilities of students. Last, we applied two adjustment methods designed to reduce bias in item parameter estimates. The proposed framework showed its potential for obtaining unbiased item parameter estimates that can then be fixed in the VLE in order to obtain ongoing ability estimates for operational purposes.

[1]NWEA, Portland, OR, USA
[2]University of Florida, Gainesville, FL, USA

Corresponding Author:
Kang Xue, NWEA, 121 NW Everett St, Portland, OR 97209, USA.
Email: kang.xue0622@gmail.com

Keywords

virtual learning environment, semisupervised learning, item response theory, missing data

Introduction

Compared with physical learning environments such as classrooms, virtual learning environments (VLEs) refer to systems that deliver learning materials to students in the digital space (Weller, 2007). VLEs allow students to engage in self-paced learning and provide the opportunity to learn new skills without having to travel. Furthermore, teachers can orchestrate learning experiences (Prieto et al., 2011) with video lessons, practice questions, and discussion forums, using VLEs. In addition, educational researchers can utilize VLE data to address pressing questions in education, such as identifying factors that affect learning (Mining, 2012). VLEs have been used extensively around the world, and have had special importance to replace in-person instruction during school closures prompted by the 2020 COVID-19 pandemic (Dhawan, 2020).

A core aspect of conducting educational research with VLE data is the challenge of measuring abilities of students who are engaging with the platform (Means & Anderson, 2013) using psychometric methods (Andrich & Luo, 1993) such as item response theory (IRT; Lord & Novick, 1968). IRT models are commonly used to estimate the psychometric properties of tests administered to students through VLEs (Park et al., 2019).

Item response data in VLEs are obtained from system logs stored in a server (Romero et al., 2014; Sheard, 2011). These data may have large percentages of missing values for many reasons, such as recording errors and students choosing to skip items. Generally, there are four types of missing data: structurally missing (i.e., data are missing for a planned purpose), missing completely at random (MCAR; i.e., the probability of missingness is the same for all cases; Rubin, 1976), missing at random (MAR; i.e., the probability of missingness is the same only within groups defined by the observed data; Rubin, 1976), and missing not at random or nonignorable missing values (MNAR; Rubin, 1987).

Several studies have applied different kinds of imputation-based approaches to deal with data classified as MCAR or MAR (e.g., Bernaards & Sijtsma, 2000; Finch, 2008; Sijtsma & Van der Ark, 2003). However, MNAR data in item responses presents a more difficult analysis challenge because the missing data may be related to the trait of measurement, which is expected to result in biased parameter estimation if ignored (Finch, 2008). The current study demonstrates that missing item response data in Math Nation, a VLE used by over 500,000 students per year (Lastinger Center for Learning, 2020), are MNAR because student decisions to skip items were partially related to their underlying ability. It also presents and evaluates a method to obtain unbiased item parameter estimates when data are MNAR, using artificial neural networks (ANNs).

Researchers have proposed various model-based approaches to reduce the impacts of MNAR item response data. In the IRT model-based approaches, the tendency to skip items is often included in the models and accounted for in the estimation of item and person parameters (Pohl et al., 2014). One type of model-based approach, the latent approach, includes modeling the tendency to have missing data via a latent missing propensity trait in a multidimensional IRT model (Holman & Glas, 2005). Another type of model-based approach, the manifest approach, involves calculating observed variables that represent missing propensity and then including those variables in measurement models (Rose et al., 2010). Studies on model-based approaches evaluated their performance in assessment data analysis when the proportion of missing data were not very high, often well under 50% of the total data in the item response matrix (e.g., 30%; Rose et al., 2010). However, the data collected in VLEs frequently contain proportions of missingness exceeding 50% because students are often self-selecting whether and when to engage with items in the VLE. This self-selection not only results in sparse data, but it also increases the chance that the missing data in the VLE are related to a wide variety of unobserved cognitive factors (e.g., academic achievement, engagement, self-regulation). Such large amounts of MNAR data do not align with the conditions under which model-based approaches to handle missingness were developed or evaluated. Therefore, there is a need for improved methods to handle large proportions of item MNAR data from system logs of VLEs.

In the past decade, due to the development of computing technology, applications of machine learning to data analysis have increased rapidly (Embretson & Reise, 2013), and there is large potential for use with missing data problems (Fazakis et al., 2020). Machine learning, which is a type of artificial intelligence, allows computer systems to learn and complete tasks without being programmed explicitly for such tasks (Bishop, 2006). Machine learning algorithms have the capacity to analyze complex and high-dimensional data, and hence applying machine learning techniques to handling missing item data in VLEs is a possible mechanism for improving psychometric research in technology-enhanced educational environments (Means & Anderson, 2013; Mining, 2012).

As one subfield of machine learning, ANNs (Goodfellow et al., 2016) have been proposed as an attractive approach to convert a pattern of item responses into latent variables (Cui et al., 2016; Cui et al., 2017; Paulsen, 2019; Xue & Bradshaw, 2021). In the current project, semisupervised learning ANNs are introduced into the IRT research area. The proposed semisupervised learning ANNs architecture may provide an accurate estimation of item parameters when item data from VLEs is MNAR. In the following sections, we first provide a brief introduction to ANN. Then, we motivate the application of ANN to MNAR data by presenting an exploratory study of the item response data collected and its patterns of missingness within a statewide database from a VLE. Next, we describe the proposed semisupervised learning framework to estimate an unbiased latent trait using deep learning techniques. After that, we present the two adjustment methods for improving the accuracy of item parameter

estimates. Then, we show the performance of the proposed framework through a simulation study. Last, we conclude with a discussion of findings, limitations, and potential future research.

Artificial Neural Networks and Semisupervised Learning

An ANN is a computational system inspired by biological neural systems for information processing in animal brains. An ANN is built on inputs being translated to outputs through a series of neuron layers. It consists of three types of layers: an input layer, a hidden layer(s), and an output layer. Each layer consists of a number of neurons (or nodes), and each node is connected to the nodes in the next layer. Each layer (except for the input layer) uses the output of its previous layer as the input. There are two advantages of using ANNs for psychometrics: first, the feature hierarchy (LeCun, 2012) supports ANNs to extract the latent variables (e.g., student ability, student attribute profiles) measured by the designed items in an assessment; second, the Universal Approximation Theorem (Csáji et al., 2001; Hornik et al., 1994; Lu et al., 2017) supports ANNs to approximate the item response equation, which represents the relationship between latent variables (e.g., student ability, student attribute profiles) and observed outcomes. These two advantages also bring ANNs promising computation techniques for tolerating some loss of quality or optimality in computed results.

Most previous research studies focused on supervised learning ANNs to classify students into different latent groups for cognitive diagnostic classification (CDM; Cui et al., 2016; Cui et al., 2017; Paulsen, 2019). In these studies, to train the supervised learning ANNs, the ideal response patterns were set as the input layer and the associated attribute profiles as the output layer. Cui et al. (2016) assumed the deterministic inputs, noisy, and gate (DINA; Junker & Sijtsma, 2001) model with both slipping and guessing parameters fixed to zero to simulate ideal responses for training a multilayer perceptron. However, item parameters are more difficult to fix when simulating from IRT model equations as they require more specific parameter values, and the results of supervised learning ANNs were not as accurate as the DINA model in the simulation study (Cui et al., 2016) for CDM. Some other research studies applied unsupervised learning ANNs (e.g., self-organizing map; Cui et al., 2016) to do latent class grouping for CDM. However, the unsupervised learning ANNs are not suitable for IRT model applications because the latent traits in IRT models are continuous rather than categorical variables.

Semisupervised learning (Zhu & Goldberg, 2009) concerns the study of how computers and natural systems learn in the presence of both labeled and unlabeled data. It is somewhere between supervised learning and unsupervised learning. The research goal of semisupervised learning is to understand how combining labeled and unlabeled data can change the machine learning behavior and allow for the design of algorithms that take advantage of such a combination. Xue and Bradshaw (2021) first provided a semisupervised learning ANN method for CDM. The semisupervised

learning architecture could refine the classification accuracy based on the initial classification obtained from two more constrained DCMs (i.e., DINA and DINO models). The experimental results showed the proposed method obtained more accurate and robust classification results than theoretical DCMs when the item response equations and Q-matrices were misspecified, and the assessment did not have a high diagnostic quality (i.e., the difference of correct response rates between the mastery group and nonmastery group is greater than 0.30). These findings inspire the notion, explored in the current study, that semisupervised learning could improve estimation accuracy in measurement applications when the data contains noise (e.g., noise resulting from nonignorable missingness). In the next section, we will motivate this application by exploring MNAR data mechanisms in a VLE.

Exploratory Study

Sample

The data consist of student responses to the ''Algebra I'' items within Math Nation, which is a VLE used statewide in Florida, as well as a few other states (Lastinger Center for Learning, 2020). This VLE has a series of instructional videos and practice items organized into 10 domains. We focused on analyzing responses to items from the entire item bank of each Algebra I domain, which are used in the VLE to randomly draw questions for 10-item quizzes. Student use of these quizzes to practice has been positively related to student achievement on Florida's high-stakes Algebra 1 End-of-Course assessment (Leite et al., 2019; Leite et al., 2021). For each domain, the number of items in the item bank ranged from 41 to 89, and we assumed that each domain consisted of a unique trait. The total number of students in the sample was 63,625, which was obtained in the 2017/2018 academic year from 71 districts in Florida (this number includes brick-and-mortar districts as well as virtual schools).

In Math Nation, students can self-select to take test items, and the VLE provides randomly selected items from the domain item bank to the student. Since students were allowed to self-select to respond to items, and to skip items presented by the VLE, the full item response matrix that crosses all 63,625 students with all items in a domain contains a large number of missing values. Figure 1 shows a visualization of the missing data, with each domain of Algebra I in Math Nation being represented by a subfigure. The x-axis of each subfigure indicates the item ID. The red, green, and blue filling refer to incorrect response, correct response, and no response (missing values), respectively. Note that for some newer items in the VLE, the missing proportions are close to 1.00, so these items were removed in the following analyses. Even removing the students who skipped one domain, the proportions of students missing responses for most items in the domain are between 0.55 and 0.75.

Generally, the response patterns of students can be classified into three categories: (1) skipped the domain (i.e., no responses to any test items within the domain), (2) completed the domain (i.e., responded to all test items within the domain), or (3) mixed response (i.e., responded to some items within the domain). We were not able

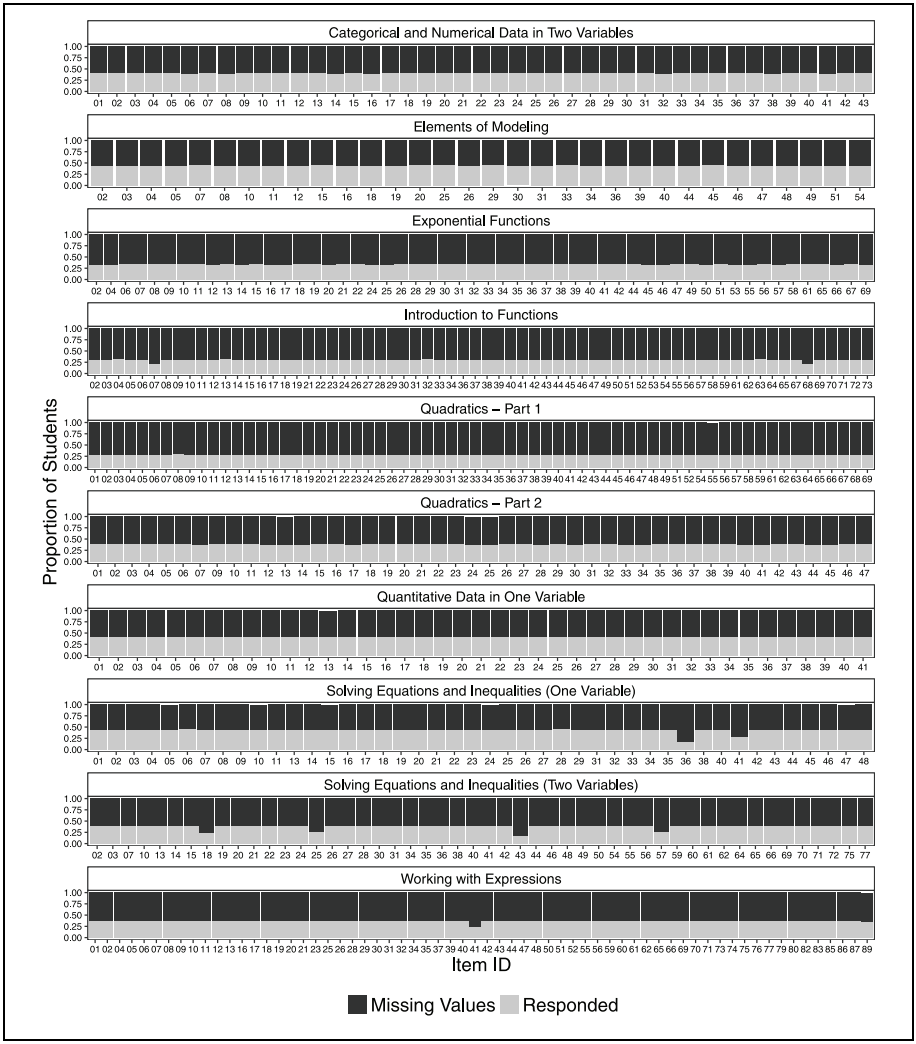

Figure 1. Visualization of missing item data in Algebra Nation Data of Florida, by domain (excluding students who did not move into the domain).

to distinguish skipped from not-reached items in our data, and hence these two forms of missing responses are grouped together under the mixed response category. Table 1 shows that although the raw frequency of students who completed the domain was relatively large, they constituted a small proportion of the total sample (less than 5%).

Table 1. Overview of the Data for Operational Data Exploration.

Domain	Number of students (63,625)			Number of items	
	Completing domain	Mixed response	Skipping domain	Total items	Available items
Working with Expressions	2,809 (4.4%)	37,675 (59.2%)	23,141 (36.4%)	89	64
Solving Equations and Inequalities with One Variable	2,749 (4.3%)	33,475 (52.6%)	27,401 (43.1%)	48	44
Solving Equations and Inequalities with Two Variables	760 (1.2%)	30,558 (48.0%)	32,307 (50.8%)	77	49
Introduction to Functions	1,749 (2.7%)	28,387 (44.6%)	33,489 (52.6%)	73	65
Quadratics—Part 1	1,983 (3.1%)	25,123 (39.5%)	36,519 (57.4%)	69	65
Quadratics—Part 2	3,008 (4.7%)	20,260 (31.8%)	40,357 (63.4%)	47	45
Exponential Functions	2,180 (3.4%)	21,864 (34.4%)	39,581 (62.2%)	69	53
Elements of Modeling	2,307 (3.6%)	16,665 (26.2%)	44,653 (70.2%)	54	32
Quantitative Data in One Variable	3,094 (4.9%)	20,822 (32.7%)	39,709 (62.4%)	41	40
Categorical and Numerical Data in Two Variables	2,761 (4.3%)	16,475 (25.9%)	44,389 (69.8%)	43	40

Note. The total number of students is 63,625. Available items were the ones with enough responses. Completing Domain is the group of students who completed all the available items in one domain; Skipping Domain was the group of students who skipped all available items in one domain; Mixed Response was the group of students who complete a part of available items in one domain.

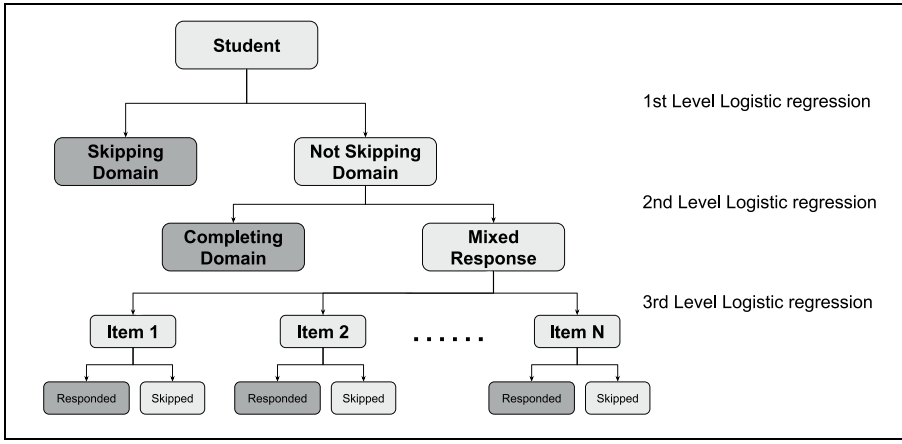

Figure 2. Flow chart of the tree-structured logistic regression.

Research Questions

In the exploratory study, we asked three questions: (1) Is there a relationship between skipping a domain and student ability? (2) Is there a relationship between not completing a domain and student ability? (3) Is there a relationship between skipping an item and student ability or item difficulty?

Analysis and Results

We used tree-structured logistic regressions (see Figure 2) to answer the questions. We did not have a latent algebra ability estimate to use in this exploratory analysis, so we used the state's standardized mathematics test scores (i.e., Florida Standards Assessment; Florida Department of Education, 2021) from the prior school year as a proxy for ability. This is the one piece of data from outside the VLE that we used in our full study to obtain our results. We also did not yet have estimates of item difficulty from an IRT model, so we used the incorrect response rates (i.e., the proportion of students who answered the item incorrectly out of all students who engaged with the item) as a proxy, mirroring a classical test theory approach. For example, if the incorrect response rate was 89%, the item difficulty was set as 0.89. The tree-structured logistic regression was applied to each domain respectively.

To address the first research question, all students were classified into two groups within each domain: students who skipped the domain and students who did not skip the domain. Within the group of students who did not skip the domain, students were grouped as either those who completed the domain or those who responded to some but not all items in the domain. Then we fit a logistic regression model to the data of each district individually as

$$P(E_i(\text{skipping a domain}) = 1) = \frac{\exp(\beta_{0,i} + \beta_{1,i}\text{ability})}{1 + \exp(\beta_{0,i} + \beta_{1,i}\text{ability})} \quad (1)$$

where i represents the ith domain. Figure 3 shows the estimates of $\beta_{1,i}$ for each district. The height of each bar indicates the number of students engaged within a domain, and the color of each bar indicates the sign of $\beta_{1,i}$ (i.e., gray is negative, dark gray is positive, and light gray is zero). From the plot, we found that for most districts and most students, $\beta_{1,i}$ were statistically significantly negative. Generally speaking, students with higher ability levels had a lower probability of skipping a domain, and students with lower ability levels had a higher probability of skipping a domain. Regarding the logistic regression model results, we could conclude that there is a relationship between skipping a domain and student ability.

To answer the second question, we divided the data set into two groups for each domain: students who completed the domain and students who did not. A logistic regression model was conducted for each district individually as

$$P(E_i(\text{not completing a domain}) = 1) = \frac{\exp(\beta_{0,i} + \beta_{1,i}\text{ability})}{1 + \exp(\beta_{0,i} + \beta_{1,i}\text{ability})} \quad (2)$$

Compared with the results of the first question, we found that estimates $\beta_{1,i}$ were not consistent across the districts (as shown in Figure 4). Taking the domain Working with Expressions as an example, the estimates of $\beta_{1,i}$ of district 06 and district 08 were significantly positive (dark gray filling), which means that within these two districts, the relationship between ability and not completing the domain Working with Expressions was positive. However, the estimates of $\beta_{1,i}$ of districts 01, 04, and 20 were statistically significantly negative (gray filling) for this domain. Thus, it was not reasonable to reach a consistent conclusion about the relationship between ability and not completing domains.

To answer the third question, in contrast to the previous two questions, the logistic regression was not applied to each district respectively for two reasons: (1) at the item level, for the students who completed a part of the domain, the district should have no impact on which particular items were shown to the student because they were randomly selected; (2) due to the high missing proportion of each item, the number of students of each district was too small to achieve a reliable estimation result in the logistic regression. With regard to the response type to the kth item in the ith domain, we put students into two groups (i.e., those who skipped the kth item and those who did not) to fit the logistic regression model defined as

$$P(E(\text{skipping an item}) = 1) = \frac{\exp(\beta_{0,ik} + \beta_{1,ik}\text{ability} + \beta_{2,ik}\text{difficulty}_k)}{1 + \exp(\beta_{0,ik} + \beta_{1,ik}\text{ability} + \beta_{2,ik}\text{difficulty}_k)} \quad (3)$$

where k indicates kth item. The estimates of $\beta_{1,ik}$ and $\beta_{2,ik}$ are shown in Table 2. For ability, the estimates of $\beta_{1,ik}$ were statistically significantly negative in all domains, indicating that a student with a higher ability level had a higher probability of

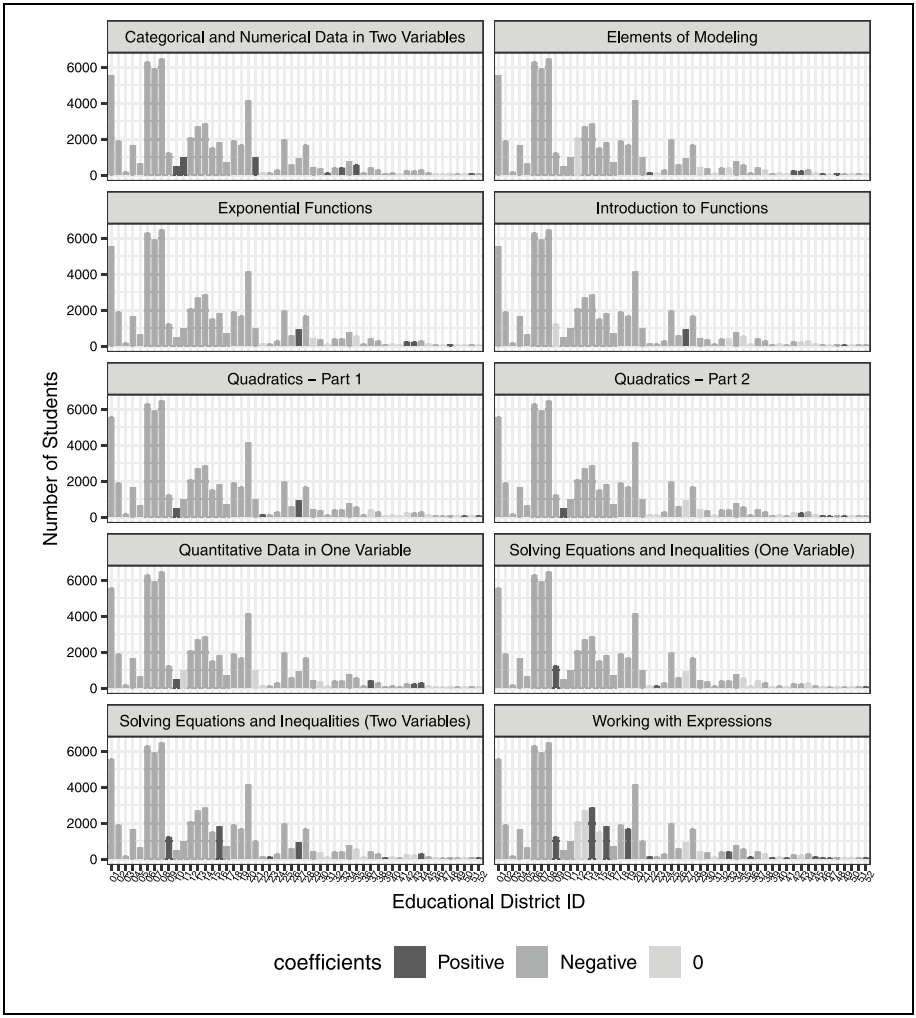

Figure 3. Results from the logistic regression analysis relating ability to skipping domains (only shows the district with over 100 students engaged).

Note. The height of each bar indicates the number of students in each district. The filling colors indicate the statistical testing results of the parameter β_1 estimation in Equation 1. Dark gray filling means β_1 was significantly positive; gray filling means β_1 was significantly negative; light gray filling means β_1 was no significant.

responding to an item shown to them. For item difficulty, the estimates of $\beta_{2,ik}$ were statistically significantly positive, indicating that skipping items was more likely as item difficulty increased. Although the VLE's random item selection within domains caused missingness in the students' response pattern, such missing data were viewed as MCAR and did not affect the significance of estimates of $\beta_{1,ik}$ and $\beta_{2,ik}$.

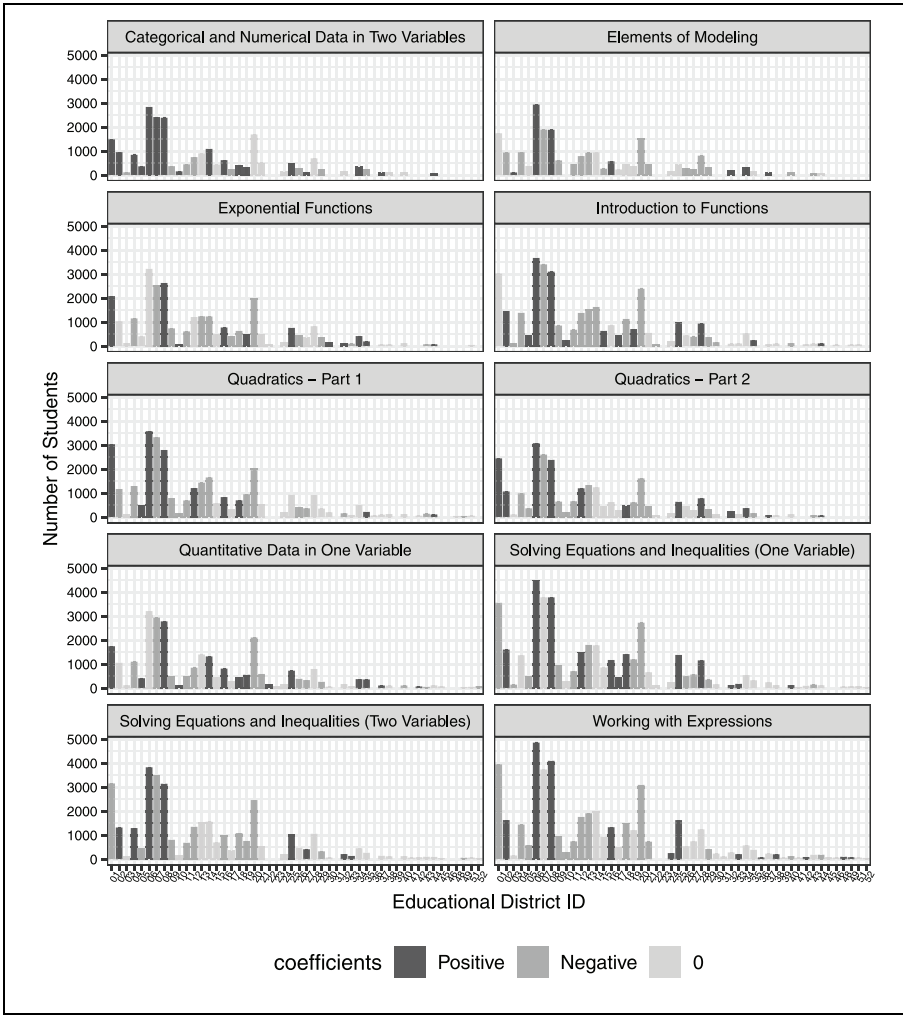

Figure 4. Results from the logistic regression analysis relating ability to not completing domain.

Note. The height of each bar indicates the number of students in each district. The filling colors indicate the statistical testing results of the parameter β_1 estimation in Equation 2. Dark gray filling means β_1 was significantly positive; gray filling means β_1 was significantly negative; light gray filling means β_1 was no significant.

Conclusion

In summary, the data exploration provided the following findings: (1) students with higher ability had a lower probability of skipping a domain; (2) students with higher ability had a lower probability of skipping items that were presented to them; (3)

Table 2. The Results of the Mixed Response Testing.

Domain	$\hat{\beta}_1$	$\hat{\beta}_1$
Working with Expressions	−0.0945 (***)	0.1221 (***)
Solving Equations and Inequalities with One Variable	−0.1800 (***)	0.1294 (***)
Solving Equations and Inequalities with Two Variables	−0.1310 (***)	0.0845 (***)
Introduction to Functions	−0.1255 (***)	0.1059 (***)
Quadratics—Part 1	−0.0592 (***)	0.1327 (***)
Quadratics—Part 2	−0.0774 (***)	0.1165 (***)
Exponential Functions	−0.0673 (***)	0.1203 (***)
Elements of Modeling	−0.0302 (***)	0.0956 (***)
Quantitative Data in One Variable	−0.0832 (***)	0.1264 (***)
Categorical and Numerical Data in Two Variables	−0.0923 (***)	0.1698 (***)

Note. $\hat{\beta}_1$ indicates the estimation of β_1 and $\hat{\beta}_2$ indicates the estimation of β_2 in Equation 3. This table shows that $\hat{\beta}_1$ was significantly negative in all 10 domains, while $\hat{\beta}_2$ was significantly positive in all 10 domains.

***p value $<$ .0001 in logistic regression test.

students were more likely to respond to easy items; and (4) choosing to not complete a domain was not consistently related to ability across the districts. We believe the final finding is related to the fact that teachers guide students through the Algebra curriculum throughout the year, and hence deciding when and which domains to use may be more teacher driven than student driven. Ultimately, we conclude that our data set not only had a large proportion of missing data for each item but also that much of that missing data was MNAR. Therefore, this result motivated the development of a semisupervised learning-based method for obtaining unbiased item parameter estimates applicable to VLE data with a large proportion of MNAR data. In the next section, we present the proposed method in detail.

A Semisupervised Learning-Based Bias Adjustment Method

The proposed semisupervised learning-based bias adjustment procedure contained two parts: (1) unbiased ability estimation using semisupervised deep learning architecture; (2) item parameter adjustment methods to obtain unbiased item parameter estimates. These two parts are discussed separately below.

Unbiased Ability Estimation Using Semisupervised Deep Learning Architecture

To apply semisupervised learning to the VLE item data from Math Nation, we first fit the two-parameter logistic (2PL) IRT model to complete data sets in each Algebra 1 domain by using only the data from students who completed all items in the domain, referred to as anchor students. However, we assume the parameter estimates

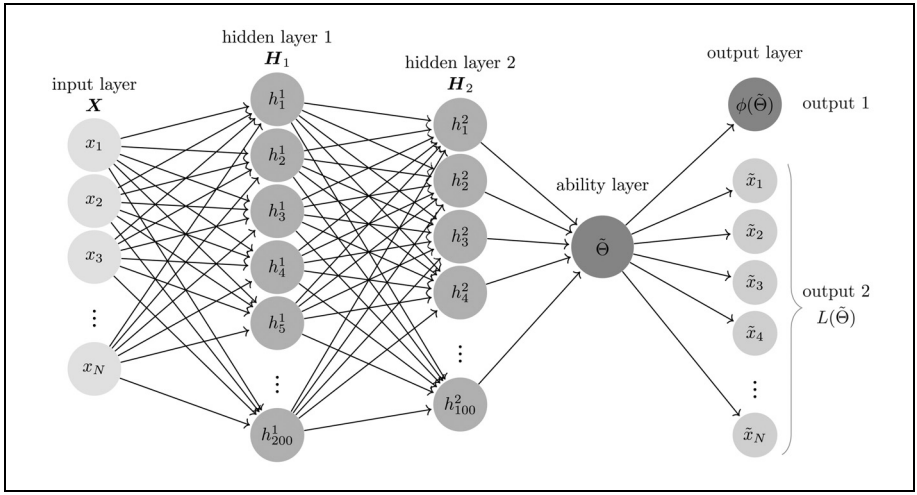

Figure 5. Results from the logistic regression analysis relating ability to completing domains (only shows the district with over 100 students engaged).

Note. A deep learning architecture with three hidden layers was used to convert the observed response patterns to the unbiased latent trait. To train the deep learning architecture, the distance between two outputs of the DFN and two targets was minimized.

are biased when obtained from only the anchor students due to the deletion of MNAR data. Thus, to achieve unbiased parameter estimation through the semisupervised learning algorithm, we needed to achieve these two goals:

1. Extract unbiased latent trait Θ for the anchor students; by using Θ, we could maximize $L(\Theta)$, the likelihood of the 2PL-IRT model
2. Specify the relationship between the biased trait estimates $\hat{\theta}$ and unbiased latent trait Θ through a function $\hat{\theta} = \phi(\Theta)$

Although a Bayesian method could be an optional method to achieve these goals, determining $\phi(\Theta)$ requires a data analysis to determine the statistical equation, and the procedure is time consuming. Considering the two advantages of using ANNs (i.e., feature hierarchy and Universal Approximation Theorem) described previously, we introduced ANNs to attain these goals. Figure 5 shows the flow chart of the proposed deep learning architecture based on deep feedforward network (DFN; Goodfellow et al., 2016), a particular ANN model that we chose because of its simplicity in architecture and training. The goals of the DFN were to convert the observed responses of anchor students to their unbiased latent traits Θ, and approximate the functions $\phi()$ and the likelihood L simultaneously.

From Figure 5, there are two hidden layers between the input layer and the latent trait layer. The number of hidden layers was based on previous research using deep

learning methods for CDM (Cui et al., 2017; Xue, 2019). Rectified linear units (ReLU; Lu et al., 2017) was chosen as the activation function to bring nonlinearity to the system. The unbiased latent trait Θ extracted using the DFN is represented as

$$\tilde{\Theta} = \Psi(X; \boldsymbol{\omega}_1) \tag{4}$$

where $\boldsymbol{\omega}_1$ is the parameters of DFN between the input layer and ability layer, and Ψ indicates the approximate computing function from the input layer to the ability layer. Given $\tilde{\Theta}$, the two outputs of the DFN structure are represented as

$$\begin{aligned} \tilde{\theta} &= \tilde{\varphi}(\tilde{\Theta}; \boldsymbol{\omega}_2) = \tilde{\varphi} \circ \Psi(X; \boldsymbol{\omega}_1) \\ \tilde{\theta} &= \tilde{L}(\tilde{\Theta}; \boldsymbol{\omega}_3) = \tilde{L} \circ \Psi(X; \boldsymbol{\omega}_1) \end{aligned} \tag{5}$$

where $\boldsymbol{\omega}_2$ and $\boldsymbol{\omega}_3$ are the parameters of the DFN connecting the ability layer and output layer. $\tilde{\varphi}$ is the approximate computing of the relationship between the biased estimation of $\tilde{\theta}$ and unbiased latent trait Θ. $\tilde{L}$ approximates the likelihood function (i.e., item response equation) of 2PL-IRT.

To train the unknown parameters $\{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \boldsymbol{\omega}_3\}$ within the DFN, we set two targets for the DFN architecture. The first target was $\hat{\theta}$, the biased students' ability estimates from fitting the 2PL-IRT model to the anchor student responses; the second target was the observed response patterns $X = \{x\}$ of the anchor students. Corresponding to the two targets, the DFN learned a function for mapping inputs to two outputs: the approximated biased ability $\tilde{\theta} = \tilde{\varphi}(\tilde{\Theta})$ and the reconstructed response pattern $\tilde{X} = \tilde{L}(\tilde{\Theta})$. The parameters of DFN, $\{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \boldsymbol{\omega}_3\}$, were estimated by minimizing the weighted cost function defined as

$$\{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \boldsymbol{\omega}_3\} \sim \arg\min\{w_1 MSE(\hat{\theta}, \tilde{\theta}) + w_2 H(X, \tilde{X})\} \tag{6}$$

In the weighted cost function, we used two kinds of error functions corresponding to two outputs. Because the first target, $\hat{\theta}$, and the first output, $\tilde{\theta}$, of DFN were continuous variables, the mean square error was used to calculate their differences. Because each variable in X was binary, cross-entropy (H) was used to calculate the differences between X and $\tilde{X}$. In Equation 6, there are two hyperparameters, w_1 and w_2, and $w_1 + w_2 = 1$. To provide a more general estimation of the parameters to avoid overfitting and increase the prediction power in a new data set, w_1 and w_2 were determined through the elbow method (Cattell, 1966; Thorndike, 1953) in the validation test. In clustering analysis, the elbow method is a heuristic used in determining the number of clusters, such as number of components in the principal component analysis, and the number of centers in K-means clustering. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use. More details about using the elbow method in validation are described in the simulation study section.

Two Item Parameter Adjustment Methods

The DFN converted observed response pattern X to unbiased ability estimation $\tilde{\Theta}$. However, the item parameter estimates from the 2PL-IRT model fitting to anchor students are still assumed to be biased and are hence in need of adjustment. We proposed and evaluated two adjustment methods to reduce bias and improve the accuracy of item parameter estimation when applying the 2PL-IRT model to the anchor student group instead of the whole population.

The first adjustment method, item equating adjustment (IEA), was inspired by the horizontal equating method (Baker, 1984), which refers to the equating of tests administered to groups with similar abilities taking similar tests. $\hat{\theta}$ and $\tilde{\Theta}$ can be viewed as the ability estimates of the same anchor student group under two tests with different item difficulties. In IEA, the ability distribution of anchor students was used as the reference population distribution. The biased item difficulty estimates were converted to unbiased item difficulty estimates with

$$\tilde{b}_j = \hat{b}_j - (\overline{\tilde{\Theta}} - \overline{\hat{\theta}}) \tag{7}$$

where $\overline{\hat{\theta}}$ and $\overline{\tilde{\Theta}}$ are the average of biased ability estimates and unbiased ability estimates respectively, $\hat{b}_j$ is the biased item difficulty estimates for jth item, and $\tilde{b}_j$ is the adjusted item difficulty estimates. IEA only reduced the biases contained in the item difficulty estimates because those estimates are on the scale of $\overline{\hat{\theta}}$, which is required for Equation 7.

The second adjustment method, bootstrapping adjustment (BA), was proposed to reduce the biases contained in both item difficulty and item discrimination parameters. The BA method includes four steps:

1. Creating a new item response set based on sampling with replacement from the anchor students considering their unbiased ability estimates $\tilde{\Theta}$ to create an ability distribution of the students in the new data set that is normally distributed, centered at zero, and that has a sample size equal to the number of anchor students
2. Apply 2PL-IRT to the new sample set to estimate the item parameters
3. Repeat Step 1 and Step 2 K times such that a set of K estimates of difficulty and discrimination of the jth item can be obtained $\{\tilde{a}_{j,k}, \tilde{b}_{j,k}\}$, where $k = \{1, ..., K\}$
4. Calculate the estimate of item discrimination as $\hat{a} = \frac{1}{K}\sum_1^K \tilde{a}_{j,k}$, and the estimate of item difficulty as $\hat{b} = \frac{1}{K}\sum_1^K \tilde{b}_{j,k}$

In our study, K was set to 100. The BA method is more robust to the biases contained in the parameter estimations because in contrast to the IEA method, BA does not assume that the distribution of ability is normal. Thus, the BA method could reduce the biases contained in both item discrimination and difficulty estimates. In addition, the BA has the potential for use with more complicated IRT models, such

as the 3PL-IRT model. However, the disadvantage of using BA is that it is more time-consuming than IEA.

Simulation Study

In this section, we report an evaluation of the proposed semisupervised learning method and adjustment methods through a simulation study with the 2PL-IRT model as the population model. The research question we addressed was: To what extent can the semisupervised learning-based method lead to a reduction in bias of item parameter estimates from the 2PL-IRT model?

Method

To imitate the operational data set used in the exploratory study, we simulated data that contained the same number of students (63,625), the same number of items in each domain, the same number of domains (10), and also the same number of districts. We used the state's standardized mathematics test scores and biased item parameters from the exploratory study as population parameters for student true ability and item parameters during data generation, with the latter obtained by fitting 2PL-IRT models to the data (treating domains as separate data sets) that contained missingness. The simulation steps for each domain were as follows:

1. We simulated the response for each item using the student pretest mathematics ability and the item parameters under the 2PL-IRT model. The item parameters were estimated by fitting the 2PL-IRT model to the observed data containing missingness.
2. Given the true ability, we calculated the probability of the student skipping the domain using the fitted logistic regression model in Equation 1 as the data-generating model. Using this probability, students were classified into a ''skipping domain'' group if $P(\text{skipping a domain}) \geq 0.5$, and ''not skipping domain'' group if $P(\text{skipping a domain}) < 0.5$. For students in the ''skipping domain'' group, their responses were set as missing values (NA).
3. For the students classified into the ''not skipping domain'' group in Step 2, we used the fitted logistic regression model in Equation 2 as the data-generating model to calculate the probability of completing domains for each student. If the probability $P(\text{not completing a domain}) < 0.5$, then the student was classified into the ''completing domain'' group (i.e., anchor students group) and we kept their responses; students with $P(\text{not completing a domain}) \geq 0.5$ were grouped into the ''not completing domain'' group.
4. For students belonging to the ''not completing domain'' group in Step 3, we used the fitted logistic regression model in Equation 3 as a data-generating model to define if a student skipped an item using student pretest ability and the observed item's incorrect response rate. For students who skipped the item

(P(skipping an item) ≥ 0.5), the response to this item was replaced by missing values.

Table 3 compares our operational data and our simulated data with respect to the average ability and the number of students for the three groups: students who skipped a domain (skipping domain group), students who completed the domain (anchor students group), and students who completed a portion of the items in a domain (mixed response group). For the anchor student group, the average ability and the number of students were very similar between the operational and simulated data. Taking the domain ''Quadratics—Part 2'' as an example, the average ability and number of students in the operational data were 0.23 and 3008, respectively, while the two values in the simulated data were 0.23 and 2964, respectively. In addition, for the skipping domain group, the average abilities in both data sets were negative, and for the mixed response group the average abilities in both data sets were positive.

Analysis and Results

We used the ''mirt'' package (Chalmers et al., 2012) in R (R Core Team, 2013) to conduct data simulation and IRT model fitting, and the ''Tensorflow'' library (Abadi et al., 2016) in python (Van Rossum & Drake, 2009) to implement the semisupervised deep learning.[1]

After simulating the data, we first fit the 2PL-IRT model to anchor students' responses for each domain to estimate the item parameters and student abilities. For example, Figure 6 shows the comparison between estimates and the true value of items contained in the domain ''Quadratics—Part 2.'' We could observe that for most items, the 95% confidence interval of the item discrimination estimates covered the true values, but the 95% confidence interval of the item difficulty estimates did not cover the true difficulty parameters. We observed similar patterns in other domains.

Then, the proposed semisupervised deep learning architecture (see Figure 5) was applied using the simulated anchor students' responses as input and using the anchor students' ability estimates and their response patterns as two targets. The unbiased abilities of anchor students were estimated by minimizing the weighted cost function in Equation 7. The validating test was conducted in the training procedure to avoid overfitting and to determine the two hyperparameters, w_1 and w_2, in Equation 6. The training set consisted of 80% anchor students, and the validation set consisted of the remaining 20% of the anchor students for early stopping, which is a simple and effective approach to avoid overtraining the ANNs. The early stopping method is used to train on the training data set but to stop training at the point when performance on a validation data set starts to degrade. The w_1 ranged from 0.05 to 0.95 with Step 0.05. Figure 7 shows two examples of determining the hyperparameters using the elbow method in the validation test for two domains. When choosing the elbow point on the blue curve (validation loss), we could choose a value, w_1. Then, regarding w_1, there is a corresponding point on the red curve (average of ability). In Figure 7, we can

Table 3. The Comparison Between Observed Operational Data and Simulated Data.

Domain	Skipping a domain		Anchor students		Mixed response	
	Observation	Simulation	Observation	Simulation	Observation	Simulation
Working with Expressions	−0.182 (23,141)	−0.0175 (16,229)	0.095 (2,809)	0.090 (2,909)	0.105 (37,675)	0.061 (44,343)
Solving Equations and Inequalities with One Variable	−0.197 (27,401)	−0.194 (20,883)	0.18 (2,749)	0.169 (2,751)	0.146 (33,475)	0.092 (39,847)
Solving Equations and Inequalities with Two Variables	−0.175 (32,307)	−0.169 (27,834)	0.240 (760)	0.203 (776)	0.179 (30,558)	0.132 (34,871)
Introduction to Functions	−0.178 (33,489)	−0.191 (28,206)	0.167 (1,749)	0.152 (1,690)	0.200 (28,387)	0.155 (33,585)
Quadratics—Part 1	−0.157 (46,519)	−0.174 (30,523)	0.146 (1,983)	0.178 (1,968)	0.216 (25,123)	0.162 (30,990)
Quadratics—Part 2	−0.143 (40,357)	−0.147 (36,571)	0.233 (3,008)	0.228 (2,964)	0.250 (20,260)	0.199 (23,946)
Exponential Functions	−0.150 (39,581)	−0.151 (35,652)	0.163 (2,180)	0.168 (2,150)	0.256 (21,864)	0.198 (25,679)
Elements of Modeling	−0.123 (44,653)	−0.133 (42,134)	0.238 (23,07)	0.218 (2,191)	0.298 (16,665)	0.270 (19,156)
Quantitative Data in One Variable	−0.141 (39,709)	−0.245 (35,554)	0.223 (3,094)	0.242 (3,059)	0.235 (20,822)	0.180 (24,868)
Categorical and Numerical Data in Two Variables	−0.119 (44,389)	−0.122 (41,103)	0.305 (2,761)	0.313 (2,755)	0.270 (16,475)	0.214 (19,623)

Note. In each cell, the first number indicates the average ability of the students; the second number in the bracket indicates the number of students.

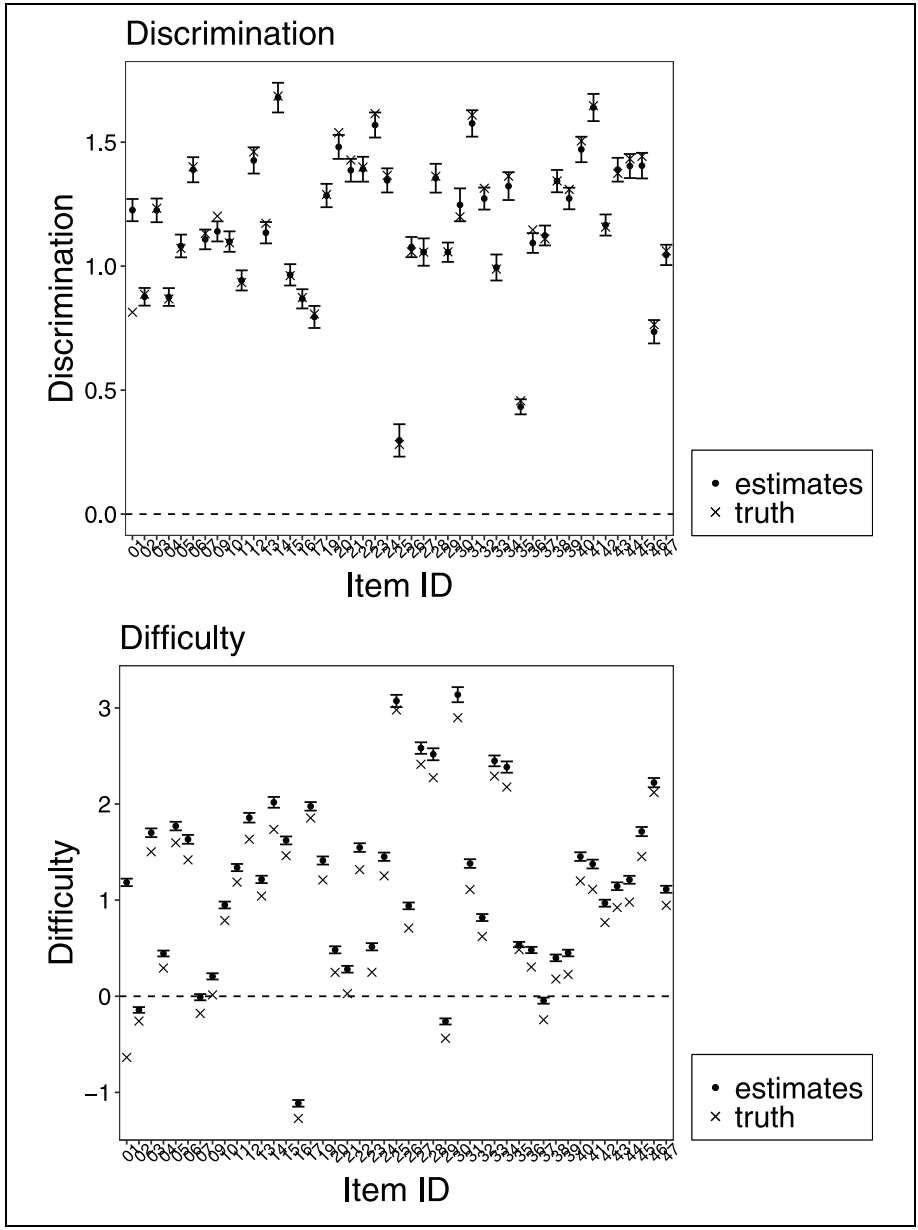

Figure 6. Comparison between estimates and the true value of items contained in the domain "Quadratics—Part 2."

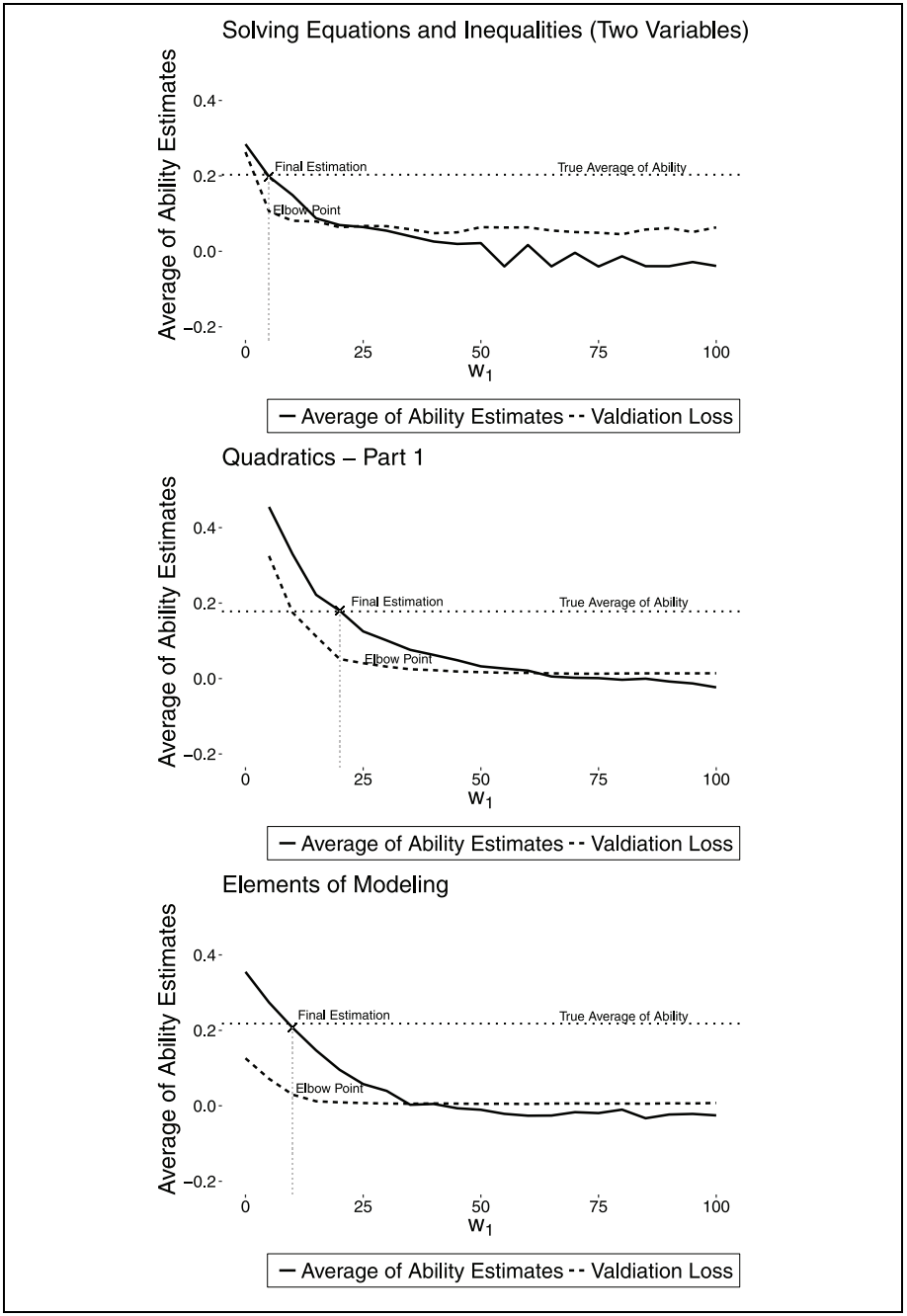

Figure 7. Three examples of choosing the appropriate hyperparameters using the elbow method.
Note. The solid curve and the dashed curve refer to the average ability estimates and the validation loss, respectively, over the weight of Target 1. The horizontal dashed line indicates the correct average ability.

Table 4. Comparison of the Distribution of Ability Estimates Between the 2PL-IRT Model Fitting to Original Data and the Proposed Semisupervised Deep Learning Architecture.

Domain	$\Theta(\sigma)$	$\hat{\theta}(\sigma)$	$\tilde{\Theta}(\sigma)$
Working with Expressions	0.090 (0.93)	−0.001 (0.99)	0.095 (0.90)
Solving Equations and Inequalities with One Variable	0.169 (0.85)	0.000 (0.98)	0.157 (0.82)
Solving Equations and Inequalities with Two Variables	0.203 (0.83)	0.000 (1.01)	0.198 (0.85)
Introduction to Functions	0.152 (0.88)	−0.001 (.99)	0.160 (0.81)
Quadratics—Part 1	0.178 (0.87)	−0.001 (1.00)	0.180 (0.88)
Quadratics—Part 2	0.228 (0.75)	−0.001 (0.99)	0.232 (0.73)
Exponential Functions	0.168 (0.85)	−0.001 (1.01)	0.171 (0.83)
Elements of Modeling	0.218 (0.79)	−0.000 (1.00)	0.207 (0.80)
Quantitative Data in One Variable	0.241 (0.77)	−0.000 (0.99)	0.241 (0.79)
Categorical and Numerical Data in Two Variables	0.312 (0.72)	−0.000 (0.98)	0.320 (0.69)

Note. In each cell, the first number refers to the average ability, and the number in the bracket refers to the standard deviation (σ) of the students' ability. Θ refers to the true ability; $\hat{\theta}$ refers to the ability estimates using direct 2PL-IRT model fitting; $\tilde{\Theta}$ refers to the ability estimates using the proposed semisupervised deep learning architecture.

observe that the value on the red curve corresponding to the elbow point on the blue curve was very close to the true average ability (the dotted line).

Because of the large number of parameters contained in the deep learning structure, the random initialization of parameters may affect the optimization when the training sample size is not large enough. Thus, one concern of using ANNs/deep learning techniques for psychometrics is that using a feature extracted by deep learning through a single training is risky as it is sensitive to the starting points of the parameters (Briggs & Circi, 2017). To solve this concern, we conducted 100 DFN trainings individually, produced ability estimates for each training, and then averaged the results as the final estimates of ability for anchor students. It took approximately 10 minutes to complete 100 DFN trainings in one domain. The total time for completing each domain was approximately 2 hours.

Table 4 compares the distribution of anchor students' ability between the biased 2PL-IRT model estimates and the proposed semisupervised deep learning architecture estimates. Table 4 shows that for each domain, the true average ability was positive and the standard deviation was less than 1. However, when fitting the 2PL model to the original data with correct, incorrect, and missing item responses (i.e., 1, 0, and NA), the average ability was consistently estimated as 0 across the domains. In contrast, the averages of the student ability estimates from the proposed semisupervised deep learning architecture were close to the true values. The proposed semisupervised deep learning architecture achieved more accurate average ability estimates without pre-data analysis to test if the sample collected could represent the population.

Parameter Estimate Bias Adjustment. After estimating the anchor students' abilities through the semisupervised deep learning architecture, the two proposed adjustment methods, IEA and BA, were conducted to reduce the biases contained in the item difficulty parameters. We chose three criteria to compare the two adjustment methods with the biased 2PL-IRT model fitting results using the anchor students' response: bias, root mean square error (RMSE), and variance of absolute bias of item difficulty parameter estimates (Variance). Bias indicates the difference between item difficulty estimates and true item difficulty parameters. RMSE indicates the distance between item difficulty estimates and true item difficulty parameters. Variance indicates the consistency of the bias adjustment across items. The bias, RMSE, and Variance were calculated as

$$\begin{aligned} bias &= \hat{b}_j - b_j \\ RMSE &= \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{b}_j - b_j)^2} \\ S^2 &= \frac{1}{J-1}\sum_{j=1}^{J}(e_j - \bar{\boldsymbol{e}}) \end{aligned} \tag{8}$$

where b_j is the true difficulty parameter for *j*th item and $\hat{b}_j$ is the estimate of that parameter. N indicates the number of items in a domain. The absolute bias for *j*th item is defined as $\hat{e}_j = |\hat{b}_j - b_j|$, and the average of the absolute biases in a domain is defined as $\bar{e} = \frac{1}{N}\sum_{j=1}^{N}\hat{e}_j$.

Figures 8, 9, and 10 compare the unadjusted item difficulty parameter estimates to the adjusted estimates based on both the IEA and BA methods. Figure 8 shows that the error results stemming from the IEA and BA method estimates cover 0 for all 10 domains, but the unadjusted item difficulty parameter estimates are biased for some domains (e.g., ''Categorical and Numerical Data in Two Variables''). Figure 9 shows that the IEA and BA methods achieved smaller RMSE results for each domain. For variance of absolute bias in Figure 10, since the IEA adjusted the difficulty estimates based on a parallel shift of the ability distribution, the variances of absolute bias from both IEA and the unadjusted estimates are equal. In other words, IEA can decrease the RMSE of parameter estimates and adjust the total bias, but the consistence of the bias adjustment is not improved. However, the BA method obtained more consistent bias adjustment (i.e., lower variance) because the bootstrapping in BA created standard-unit normally distributed samples, which matched the assumption of the original IRT estimation.

Discussion and Conclusion

With the proliferation of VLE usage in educational systems (King & South, 2017), there is a desire to obtain accurate estimates of student ability from VLE log file data,

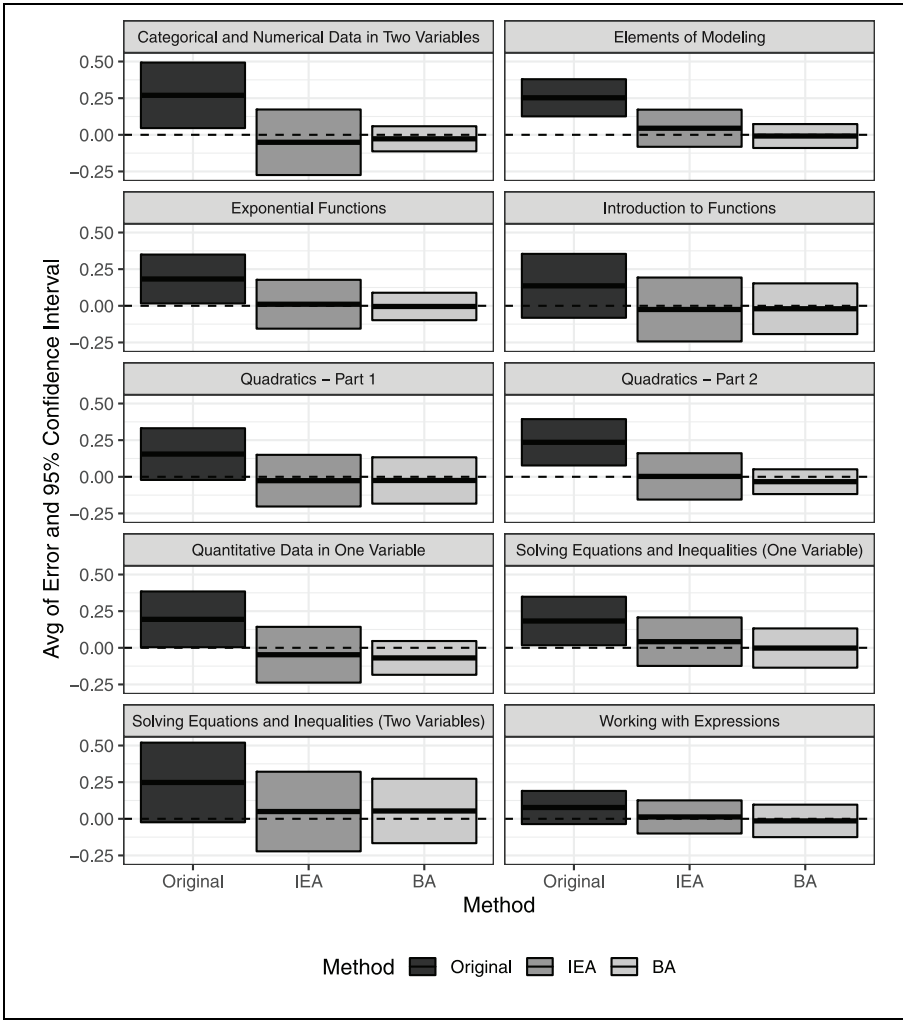

Figure 8. Bias comparison of item difficulty estimates across the unadjusted results (Original), the item equating adjustment (IEA) results, and the bootstrapping adjustment (BA) results.

which would most often needs to be done through item response data. This is complicated by the fact that the unstructured item response data often contains a large proportion of missingness. However, if one can obtain unbiased IRT item parameter estimates associated with the item response data, fixed item parameters can be used in a variety of ways to obtain accurate ability estimates for students who have taken different sets of items in the VLE.

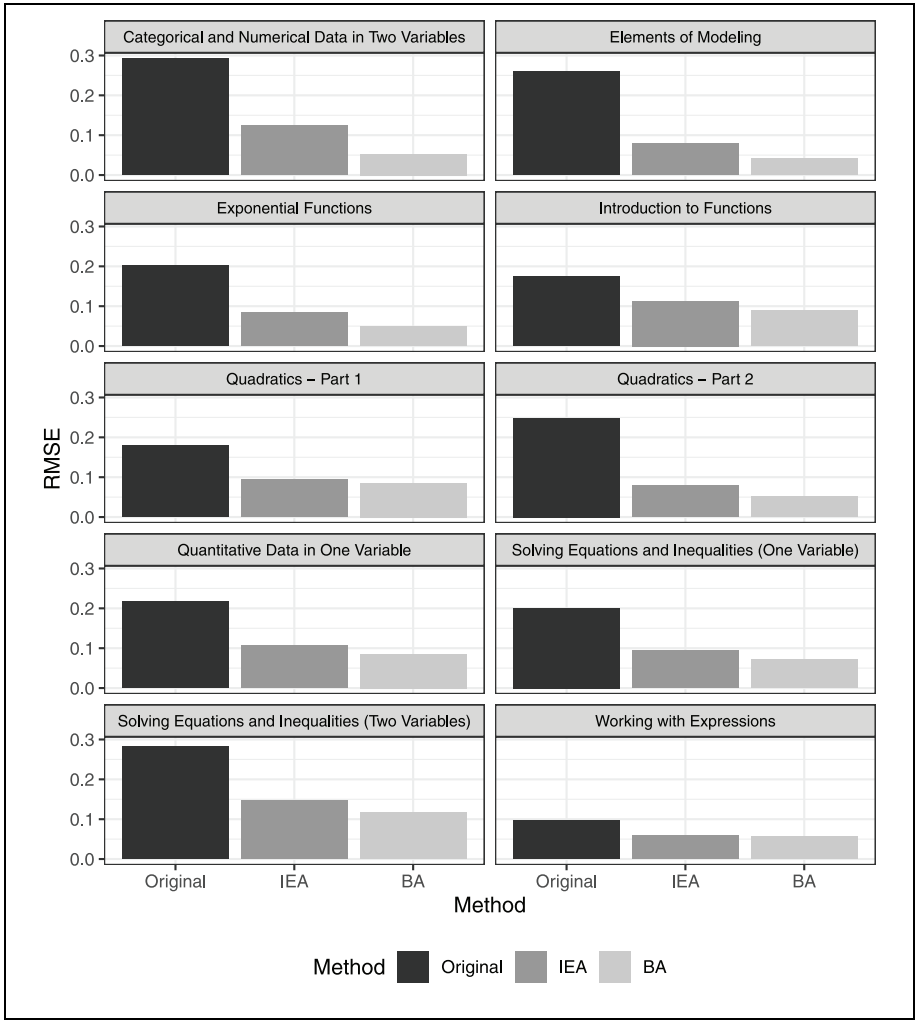

Figure 9. RMSE comparison of the item difficulty estimates across the unadjusted results (Original), the item equating adjustment (IEA) results, and the bootstrapping adjustment (BA) results.

To tackle the challenge of obtaining unbiased item parameter estimates from sparse VLE item response data, we first explored a statewide-used VLE data set to test hypotheses surrounding the relationships between the missing data and student ability and item difficulty. The exploratory study showed that the missingness was nonignorable (i.e., MNAR), which indicated that ignoring it would affect parameter estimation accuracy. Restated, the estimated parameters are unbiased for the students

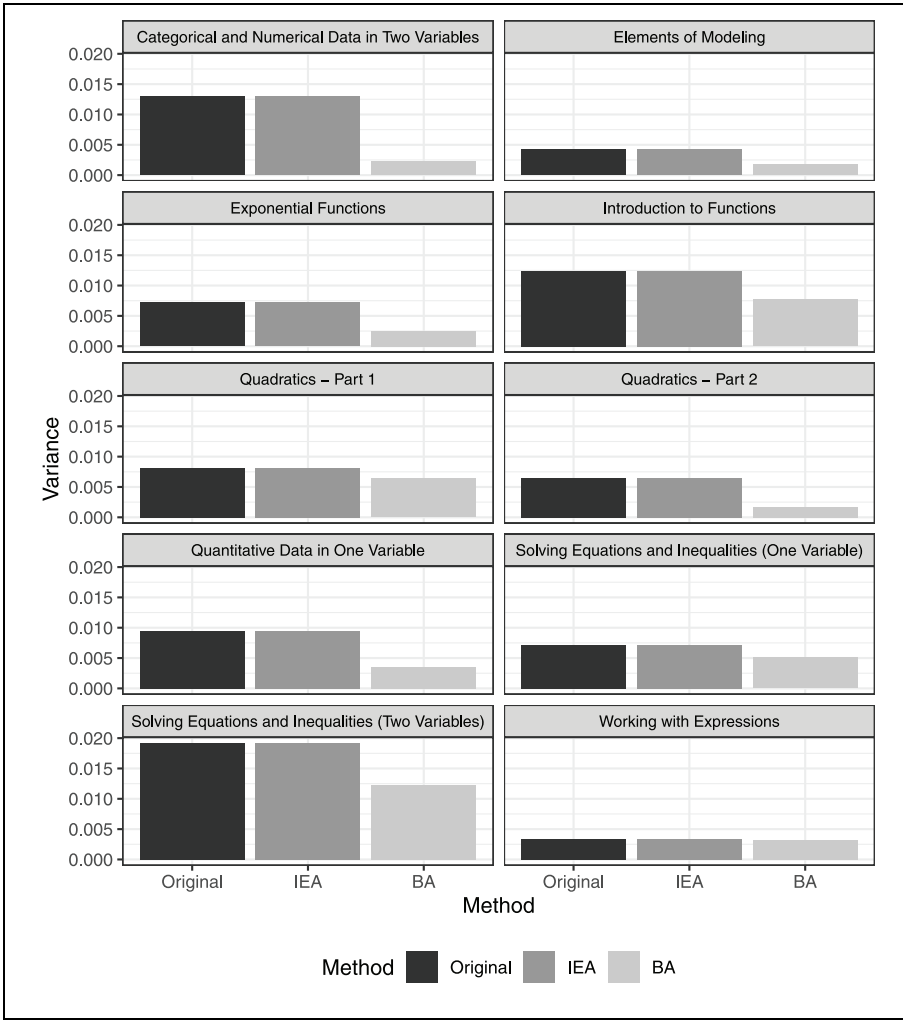

Figure 10. Variance of absolute bias comparison of the item difficulty estimates across the unadjusted results (Original), the item equating adjustment (IEA) results, and the bootstrapping adjustment (BA) results.

who completed all items (i.e., anchor students in this research) but might be biased for the students who did not complete all items in the VLE (i.e., students skipped some domains, students skipped some items within domains, or students did both).

To solve this problem, we proposed a semisupervised learning framework that is novel to IRT applications. In the proposed framework, we used deep learning to convert observed response patterns to continuous latent traits and approximate some continuous functions, which are otherwise hard to simultaneously specify

mathematically. In addition, using the ability estimates through the semisupervised learning framework, we also proposed two kinds of adjustment methods to improve the item parameter estimation accuracy. The simulation results showed that the proposed framework adjusted the biases contained in both students' ability estimates and item parameter estimates under the 2PL-IRT model. We believe that many researchers using VLE data can benefit from following our proposed method in order to obtain critical, accurate information about item parameters that can be used to support ongoing ability estimation in the VLE platform.

There were still some limitations in the proposed framework. The first limitation is that some anchor students (students who did not skip items in a domain) are required for the proposed method. In the absence of a sufficient number of anchor students, the deep learning architecture could potentially be modified to a multiclass classification model that considers the missing values as the third response type among the correct and incorrect responses. One potential advantage of this combination is that all students' responses (including missingness) could be used as training data. The second limitation is that a large data set is required for training in the deep learning architecture, but for new items in a VLE platform, the sample size of responses may not be large enough. Transfer learning (Pan & Yang, 2009) and item characteristic prediction using natural language processing (Manning & Schütze, 1999) are potentially applicable solutions to this problem. Transfer learning and natural language processing could provide better starting values and initialization of the parameters for training models on new items with few responses from students.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, 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 the views of the Institute or the U.S. Department of Education.

ORCID iDs

Kang Xue https://orcid.org/0000-0003-2161-6931
Walter Leite https://orcid.org/0000-0001-7655-5668

Note

1. Example R and Python code demonstrating the proposed methods are available at the Open Science Framework website (https://osf.io/2t5vg/?view_only=2bd0997067a1414a9dc89264d8efc0fc)

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P.,. . .Zheng, X. (2016, November 2-4). *Tensorflow: A system for large-scale machine learning* [Paper presentation]. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI, 16), Savannah, GA. https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, *17*(3), 253-276. https://doi.org/10.1177/014662169301700307

Baker, F. B. (1984). Ability metric transformations involved in vertical equating under item response theory. *Applied Psychological Measurement*, *8*(3), 261-271. https://doi.org/10.1177/014662168400800302

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research*, *35*(3), 321-364. https://doi.org/10.1207/S15327906MBR3503_03

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Briggs, D. C., & Circi, R. (2017). Challenges to the use of artificial neural networks for diagnostic classifications with student test data. *International Journal of Testing*, *17*(4), 302-321. https://doi.org/10.1080/15305058.2017.1297816

Cattell, R. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 629-637. https://doi.org/10.1207/s15327906mbr0102_10

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

Csáji, B. C. (2001). *Approximation with artificial neural networks* [Unpublished doctoral dissertation]. Etvs Lornd University.

Cui, Y., Gierl, M., & Guo, Q. (2016). Statistical classification for cognitive diagnostic assessment: An artificial neural network approach. *Educational Psychology*, *36*(6), 1065-1082. https://doi.org/10.1080/01443410.2015.1062078

Cui, Y., Guo, Q., & Cutumisu, M. (2017, June 25-28). *A neural network approach to estimate student skill mastery in cognitive diagnostic assessments* [Conference session]. 10th International Educational Data Mining Conference, Wuhan, China. https://doi.org/10.7939/R35H7C71D

Dhawan, S. (2020). Online learning: A panacea in the time of covid-19 crisis. *Journal of Educational Technology Systems*, *49*(1), 5-22. https://doi.org/10.1177/0047239520934018

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Fazakis, N., Kostopoulos, G., Kotsiantis, S., & Mporas, I. (2020). Iterative robust semi-supervised missing data imputation. *IEEE Access*, *8*, 90555-90569. https://doi.org/10.1109/ACCESS.2020.2994033

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225-245. https://doi.org/10.1111/j.1745-3984.2008.00062.x

Florida Department of Education. (2021). *Florida Standards Assessment*. http://www.fldoe.org/accountability/assessments/k-12-student-assessment/fsa.stml

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Holman, R., & Glas, C. A. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 1-17. https://doi.org/10.1111/j.2044-8317.2005.tb00312.x

Hornik, K., Stinchcombe, M., White, H., & Auer, P. (1994). Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. *Neural Computation*, *6*(6), 1262-1275. https://doi.org/10.1162/neco.1994.6.6.1262

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

King, J., & South, J. (2017). *Reimagining the role of technology in higher education: A supplement to the national education technology plan*. US Department of Education, Office of Educational Technology.

Lastinger Center for Learning. (2020). *Math nation*. https://lastinger.center.ufl.edu/mathematics/math-nation/

LeCun, Y. (2012). Learning invariant feature hierarchies. In A. Fusiello, V. Murino, & R. Cucchiara (Eds.), *Computer Vision: ECCV 2012. Workshops and demonstrations: ECCV 2012*. Lecture Notes in Computer Science (*Vol. 7583*). Springer. https://doi.org/10.1007/978-3-642-33863-2_51

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

Leite, W. L., Jing, X., Kuang, H., Kim, D., & Huggins-Manley, A. C. (2021). Multilevel mixture modeling with propensity score weights for quasi-experimental evaluation of virtual learning environments. *Structural Equation Modeling*. Advance online publication. https://doi.org/10.1080/10705511.2021.1919895

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lu, Z., Pu, H., Wang, F., Hu, Z., & Wang, L. (2017). The expressive power of neural networks: A view from the width. In M. I. Jordan, Y. LeCun, & S. A. Solla (Eds.), *Advances in neural information processing systems* (pp. 6231-6239). MIT Press.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Means, B., & Anderson, K. (2013). *Expanding evidence approaches for learning in a digital world*. Office of Educational Technology, U.S. Department of Education.

Mining, T. E. D. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief*. U.S. Department of Education, Office of Educational Technology.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345-1359. https://doi.org/10.1109/TKDE.2009.191

Park, J. Y., Cornillie, F., van der Maas, H. L., & Van Den Noortgate, W. (2019). A multidimensional IRT approach for dynamically monitoring ability growth in computerized practice environments. *Frontiers in Psychology*, *10*, Article 620. https://doi.org/10.3389/fpsyg.2019.00620

Paulsen, J. (2019). *Examining cognitive diagnostic modeling in small sample contexts* [Unpublished doctoral dissertation]. Indiana University.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item

response theory models. *Educational and Psychological Measurement*, *74*(3), 423-452. https://doi.org/10.1177/0013164413504926

Prieto, L. P., Holenko Dlab, M., Gutiérrez, I., Abdulwahed, M., & Balid, W. (2011). Orchestrating technology enhanced learning: A literature review and a conceptual framework. *International Journal of Technology Enhanced Learning*, *3*(6), 583-598. https://doi.org/10.1504/IJTEL.2011.045449

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Romero, C., Romero, J. R., & Ventura, S. (2014). A survey on pre-processing educational data. In *Educational data mining* (pp. 29-64). Springer.

Rose, N., Von Davier, M., & Xu, X. (2010). Modeling nonignorable missing data with item response theory (IRT). *ETS Research Report Series*, *2010*(1), i-53. https://doi.org/10.1002/j.2333-8504.2010.tb02218.x

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, *82*(398), 543-546. https://doi.org/10.2307/2289460

Sheard, J. (2011). Basics of statistical analysis of interactions data from web-based learning environments. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. d Baker (Eds.), *Handbook of educational data mining* (pp. 27-40). CRC Press.

Sijtsma, K., & Van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research*, *38*(4), 505-528. https://doi.org/10.1207/s15327906mbr3804_4

Thorndike, R. L. (1953). Who belongs in the family. *Psychometrika*, *18*, 267-276. https://doi.org/10.1007/BF02289263

Van Rossum, G., & Drake, F. L. (2009). *Python 2.6 reference manual* (*Python documentation manual Part 2)*. CreateSpace Independent.

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

Xue, K. (2019, August). Computational diagnostic classification model using deep feedforward network based semi-supervised learning [Paper presentation]. In 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) Workshop on Deep Learning for Education, Anchorage, AK.

Xue, K., & Bradshaw, L. P. (2021). A semi-supervised learning-based diagnostic classification method using artificial neural networks. *Frontiers in Psychology*, *11*, Article 3992. https://doi.org/10.3389/fpsyg.2020.618336

Zhu, X., & Goldberg, A. B. (2009). *Introduction to semi-supervised learning* (Synthesis Lectures on Artificial Intelligence and Machine Learning). https://doi.org/10.2200/S00196ED1V01Y200906AIM006