Semi-supervised Learning Method for Adjusting Biased Item Difficulty Estimates Caused by Nonignorable Missingness under 2PL-IRT Model [*]

Kang Xue
NWEA & University of Georgia
kang.xue@nwea.org,
kangxue@uga.edu

Anne Corinne
Huggins-Manley
University of Florida
amanley@coe.ufl.edu

Walter Leite
University of Florida
Walter.Leite@coe.ufl.edu

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 impact IRT item parameter estimation accuracy, which then negatively impacts any future ability estimates utilized in the VLE. In the psychometric literature, methods for handling missing data are mostly centered around conditions in which the data and the amount of missing data are not as large as those that come from VLEs. In this paper, we introduce a semi-supervised learning method to deal with a large proportion of missingness contained in VLE data from which one needs to obtain unbiased 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.

Keywords

virtual learning environment, semi-supervised learning, item response theory, missing data

1. INTRODUCTION

In contrast to physical learning environments such as classrooms, a virtual learning environment (VLE) refers to a system that delivers learning materials to students in a digital space. Item response theory (IRT) [3] refers to a family of mathematical models that attempt to explain the relationship between latent traits (unobservable skills or knowledge) and their manifestations (i.e. observed outcomes, responses or performance) using different statistic functions (e.g. Rasch Model, 2PL-IRT, multidimensional IRT). To estimate the item parameters for further personal adaptive learning (e.g., providing appropriate item which matches student's ability could encourage student to complete it), IRT models are widely used to determine the psychometric properties of items through analyzing students' responses in VLE [9].

How to reduce the impact of missing values on item parameter estimation of IRT models is a very common issue for data analysis and attracts lots of research attention. Generally, missing values could be categorized to 4 classes: structurally missing data, missing completely at random (MCAR), missing at random (MAR) and missing not at random (i.e. nonignorable missing values) [12]. In contrast to other types of missing values, nonignorable missing values in assessment are more complicated because they are usually caused by latent factors to be measured by IRT models. For assessment data, researchers has proposed different model-based approaches to reduce the impacts from nonignorable missing values [10]. One model-based approach, the latent approach, includes missing tendency via a latent missing propensity that is accounted for in a multidimensional IRT model [4]; another model-based approach, the manifest approach, includes missing tendency by modeling a manifest missing variable that is accounted for in a unidimensional missingness propensity [11].

However, in contrast to assessment, the data collected in VLE often contain large proportion of missingness when students are allowed to skip questions in some online courses. It makes that the missing data in VLEs are caused by a variety of cognitive and motivational factors (e.g., excess challenge, lack of challenge or lack of time). The model-based approaches are not suitable to deal with such kinds of missingness in the data collected from VLE, because determining the latent missing propensity will be very complicated for drawing inferences to model the joint distribution of the missingness and the item responses [6].

The technological changes across learning, instruction and assessment start to bring machine learning techniques into psychometrics because machine learning algorithms have the

[*]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.

Kang Xue "Semi-supervised Learning Method for Adjusting Biased Item Difficulty Estimates Caused by Nonignorable Missingness under 2PL-IRT Model" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 715 - 719

capability to analyzing complex and high-dimensional data. Applying data mining and machine learning techniques to VLE data is a mechanism to improve research in technology-enhanced educational environments [1, 8]. For example, IRT psychometric models are usually based upon logistic regression techniques which are used to be popular in solving classification problem in machine learning [7]. As a sub-field of machine learning, the primary goal of deep learning is to extract the latent variables from the input distribution using artificial neural networks (ANNs) which is a computational system inspired by biological neural networks [5]. In educational research area, deep learning has been applied for different tasks, such as automatic item generation (AIG) [14], automated scoring [13], and item characteristics prediction [17].

Inspired by the research using deep learning and semi-supervised learning techniques for cognitive diagnostic classification [15], we proposed a semi-supervised deep learning framework to reduce the impact on item parameter estimation caused by nonignorable missing values when applying two-parameter IRT (2PL-IRT) model to the data collected in VLE. The research in this paper consists of two parts: (1) exploring the real data collected within a statewide-used VLE to test if the missingness was caused by student ability and item difficulty which were measured in 2PL-IRT; and (2) proposing a semi-supervised learning method using deep learning techniques to adjust the bias in estimation caused by missingness. In the following part of this paper, we will firstly introduce the operational data exploration on the data collected within a VLE; then the semi-supervised learning method will be described in detail; the simulated study shows the performance of the proposed framework in dealing with nonignorable missingness; lastly, we will conclude the findings and limits in this framework and discuss some potential future research.

2. OPERATIONAL DATA EXPLORATION

The data collected in this research were students' responses to the "Algebra I" items within a statewide-used VLE system. The dataset contains 10 algebra domains, and we treated each domain as having its own ability to measure. The number of items ranged from 41 to 89 across domains. The total number of students was 63,625. Since students were allowed to skip items in the learning environment when they responded to the items which were selected by the system randomly, the responses to each item contained large amount of missing values. The proportion of missingness for each item is between 55% to 75%. Generally, the response patterns of students could be classified into 3 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), 3) mixed response (i.e., responded to some items within the domain).

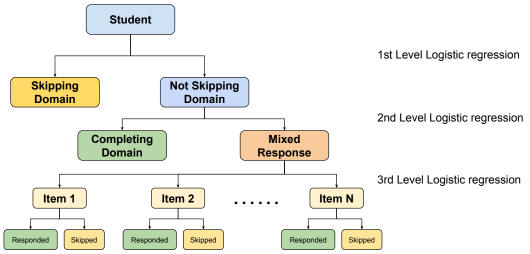

Figure 1: The diagram of the hierarchical logistic regression.

To test if the missingness was related to the item and person parameters in the 2PL-IRT model, a hierarchical logistic regression (Figure 1) was conducted for each domain individually. The hierarchical logistic regression was consisted of (1) **skipping domain test** was to test if skipping a domain related to the students' ability; (2) **completing domain test** was to test if completing a domain related to the students' ability; (3) **mixed response test** was to test if student skipping an item related to the item difficulty and students ability. As an area of mathematics, there is high correlation between the math skills and algebra skills. Thus, we used the pretest mathematical scores on the state standardized test, S, as student's true ability for the data exploration. To evaluate the relationship between ability and skipping a domain, all the students' responses were classified to two groups: students skipped the domain and students didn't skip the domain. The second group contained students completed the domain and students with mixed responses. Then the logistic regression test was conducted for each school district individually as following:

$$logit(\text{skipping domain}) = \beta_{0,ij} + \beta_{1,ij} S \tag{1}$$

where j indicates the jth educational district and i refers to the ith domain. After fitting the models, we found that for most school districts and most students, $\beta_{1,ij}$ were significant negative. We can conclude that students with high ability level had a lower probability to skip a domain, and students with low ability level had higher probability to skip a domain.

After doing skipping domain test, in the completing domain test, the dataset only contained students who didn't skip the domain. The dataset was divided to two groups: students completed domain and students with mixed response. The logistic regression test was conducted for each school district individually as following:

$$logit(\text{completing domain}) = \beta_{0,ij} + \beta_{1,ij} S \tag{2}$$

where j indicates the jth educational district and i refers to the ith domain. In contrast to the observation of "skipping domain", it was not reasonable to reach a consistent conclusion about the relationship between the ability and completing domain.

In the last subtest, two factors, students' ability and item difficulty, were assumed to impact the probability that a student responded to an item. We chose the observed incorrect response rate of the item, D_k, to indicate the item difficulty. The logistic regression was as following:

$$logit(\text{skipping } k\text{th item}) = \beta_{0,ik} + \beta_{1,ik} S + \beta_{2,ik} D_k \tag{3}$$

where i is the ith domain and k indicates the kth item. The logistic regression test showed that students with lower ability level had higher probability to skip an item shown to them; and student had a higher probability to skip item with higher difficulties. From the data exploration, we could conclude that the missing values in the data collected contained nonignorable missingness because they were caused the factors have relationship with the latent variables measured in the 2PL-IRT model.

3. SEMI-SUPERVISED DEEP LEARNING-BASED BIAS ADJUSTMENT

Intuitively, there was no missing value in the response from the anchor students, who completed all items in a domain. However, directly applying 2PL-IRT model to the anchor students would impact a parameter invariance because from the data exploration also showed there existed difference between the sub-population of anchor students and whole population. To adjust the biased ability estimates and item parameters estimates through directly applying 2PL-IRT model to the anchor students, we proposed a semi-supervised deep learning-based bias adjustment procedure which consisted of the unbiased ability estimation through a semi-supervised deep learning architecture, and the item parameter adjustment methods.

3.1 Semi-supervised deep learning architecture

The thinking of semi-supervised learning was used to improve the robustness of binary latent person variables (e.g. attribute mastery status) estimation [15]. In this research, because the latent person variables measured in 2PL-IRT model were continuous, the semi-supervised learning techniques were conducted based on the following two assumptions:

1. Given the unbiased latent trait Θ for each student, the biased estimation $\hat{\theta}$ directly using 2PL-IRT could be represented through a function: $\hat{\theta} = \phi(\Theta)$;

2. The unbiased latent trait Θ could maximize the likelihood function $P(X = 1; \Theta) = L(\Theta)$ which indicates the relationship between latent trait Θ and item response pattern $\mathbf{X} = \{x\}$;

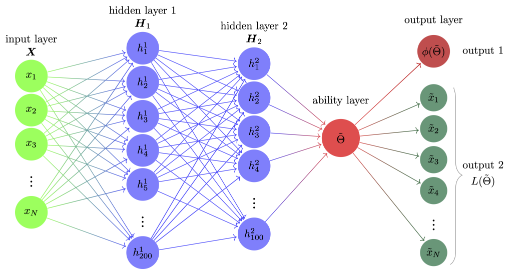

Figure 2: The diagram of the proposed semi-supervised deep learning architecture for unbiased ability estimation. In this framework, a deep learning architecture with 3 hidden layers was used to convert the observed response patterns to the unbiased ability. To train the deep learning architecture, the distance between two outputs of the DFN and two targets was minimized.

Regarding to these assumptions, the goals of the proposed semi-supervised deep learning structure to extract the unbiased latent trait Θ from the anchor students response data and approximate the function $\phi(\Theta)$ which indicated the relationship between unbiased latent trait Θ and biased estimation $\hat{\theta}$ and $L(\Theta)$ which indicates the relationship between latent trait Θ and item response pattern $\mathbf{X} = \{x\}$. From Figure 2, there were three hidden layers between the input layer and the latent trait layer. The number of hidden layers were set based on the previous research of using deep learning method for cognitive diagnostic models [16, 2]. To bring the nonlinearity to the DFN, Rectified Linear Units (ReLU) was chosen as the activation function. The unbiased latent trait $\tilde{\Theta}$ extracted using the DFN could be represented as: $\tilde{\Theta} = \Phi(\mathbf{X}; \boldsymbol{\omega})$. $\boldsymbol{\omega}$ were the connection weights in the DFN. The parameters of DFN, $\boldsymbol{\omega}$, were estimated by minimizing the following weighted cost function:

$$\boldsymbol{\omega} = \arg\min(w_1 MSE(\hat{\theta}, \phi(\tilde{\Theta})) + w_2 H(\tilde{\mathbf{X}}, \mathbf{X})) \tag{4}$$

where $\hat{\theta}$ is the biased students' ability estimation directly fitting 2PL-IRT model to the anchor students' responses; $\mathbf{X} = \{x\}$ is the observed response patterns of the anchor students. In the weighted cost function, we used two kinds of error functions corresponding to two outputs respectively: the mean square error (MSE) was used to calculate the difference between continuous variables $\hat{\theta}$ and $\phi(\tilde{\Theta})$; the cross-entropy (H) was used to calculate the difference between binary variables $\mathbf{X}$ and $\tilde{\mathbf{X}}$. The two hyperparameters, w_1 and w_2, were determined using the elbow method in validation test.

3.2 Two item parameter adjustment methods

After obtaining the parameter estimation through the training procedure, the DFN converted observed response pattern $\mathbf{X}$ to unbiased ability estimation $\tilde{\Theta}$. To reduce the biases contained in the item difficulty, two kinds of adjustment methods, item equating adjustment (IEA) and bootstrapping adjustment (BA), were proposed using the unbiased ability estimation $\tilde{\Theta}$.

IEA was inspired by the common group equating design in IRT. In IEA, the ability distribution of anchor students was the frame of reference. Then the biased item difficulty estimates were placed onto unbiased item difficulty via $\tilde{b}_j = \hat{b}_j - (\bar{\tilde{\Theta}} - \bar{\hat{\theta}})$. $\bar{\hat{\theta}}$ and $\bar{\tilde{\Theta}}$ 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 it held an assumption that the item discrimination estimates were not biased.

In contrast to IEA, BA was proposed to reduce the biases contained in both item difficulty and item discrimination parameters using bootstrapping in statistics. There were 4 steps contained in BA method:

1. Randomly sampled from the anchor students based on the unbiased ability estimates $\tilde{\Theta}$ to make the ability distribution of the new sample set is standard normal distribution and the sample size was same as the original anchor students;

2. Apply 2PL-IRT to the new sample set and estimate the item difficulty parameters and item discriminating parameters;

3. Repeated step 1 and step 2 K times, a group of estimates of difficulty and discriminating of jth item could be obtained $\{\tilde{a}_{j,k}, \tilde{b}_{j,k}\}$, where $k = 1, K$;

Table 1: Comparison of the distribution of ability estimates between directly 2PL-IRT model fitting ($\hat{\theta}$) and the proposed semi-supervised deep learning architecture ($\tilde{\Theta}$).

Domains	True $\Theta(\sigma)$	$\hat{\theta}(\sigma)$	$\tilde{\Theta}(\sigma)$
1	0.090 (0.93)	-0.001 (0.99)	0.095 (0.90)
2	0.169 (0.85)	0.000 (0.98)	0.157 (0.82)
3	0.203 (0.83)	0.000 (1.01)	0.198 (0.85)
4	0.152 (0.88)	-0.001 (0.99)	0.160 (0.81)
5	0.178 (0.87)	-0.001 (1.00)	0.180 (0.88)
6	0.228 (0.75)	-0.001 (0.99)	0.232 (0.73)
7	0.168 (0.85)	-0.001 (1.01)	0.171 (0.83)
8	0.218 (0.79)	-0.000 (1.00)	0.207 (0.80)
9	0.241 (0.77)	-0.000 (1.00)	0.241 (0.79)
9	0.312 (0.72)	-0.000 (0.98)	0.320 (0.69)

4. Then the estimate of item discrimination equaled to $\frac{1}{K}\sum_{1}^{K}\tilde{a}_{j,k}$, and the estimate of item difficulty equaled to $\frac{1}{K}\sum_{1}^{K}\tilde{b}_{j,k}$.

The BA method relies on less constraint and could reduce the biases contained in both item discrimination and difficulty estimates. The BA has the potential for applying on more complicated IRT models, such as 3PL-IRT.

4. SIMULATED STUDY

The proposed methods were tested through a simulation study under 2PL-IRT model. In the simulated study, we used "mirt" package in R to conduct data simulation and IRT model fitting and used "Tensorflow" toolbox in python to achieve the unbiased ability estimates through the semi-supervised deep learning architecture. To create data under 2PL-IRT, the known pretest mathematical ability were used as the students' ability, and the biased item parameters obtained through directly applying 2PL-IRT to the anchor students were used as item parameters. The fitted functions 1, 2, and 3 in data exploration were used to predict the students' response patterns (e.g., skipping domain, completing domain, mixed response). We selected the response of anchor students who completed all items in a domain as the input of our proposed method.

First, we applied the 2PL-IRT model directly to the simulated anchor students' responses for each domain to estimate the item parameters and students' ability. Then, the proposed semi-supervised deep learning architecture was applied using the simulated anchor students' responses as input and using the anchor students' ability estimates and their response patterns as two targets. By minimizing the weighted cost function in Equation 4, the unbiased ability of anchor students was estimated. The validating test was conducted in the training procedure to avoid over-fitting and determine the two hyperparameters w_1 and w_2 in Equation 4. Table 4 compares the distribution of ability estimates be- tween directly 2PL-IRT model fitting and the proposed semi- supervised deep learning architecture.

Using the estimation of the anchor students' ability through the semi-supervised 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 two criteria, rooted mean squared error (RMSE) and variance, to evaluate the bias adjustment methods. RMSE indicates the distance between item difficulty estimates and true item difficulty parameters, and the variance indicates the consistency of the estimates from different methods. From Figure 3, in contrast to the directly 2PL-IRT model fitting, both IEA and BA achieved much less RMSE for each domain. For variance, since the IEA adjusted the difficulty estimates based on a parallel shift of the ability distribution, the variance of IEA and directly 2PL-IRT results were the same. However, the BA method obtained more consistent estimates because bootstrapping in BA created standard normal distributed samples which matched the assumption of original IRT estimation. From the experimental results, both IEA and BA had the ability to adjust the biases contained in the estimates of item difficulty using directly 2PL-IRT model fitting. Compared with IEA which only reduce the biases of item difficulty parameters, BA method had the potential to reduce the biases contained in the item parameters for different IRT models.

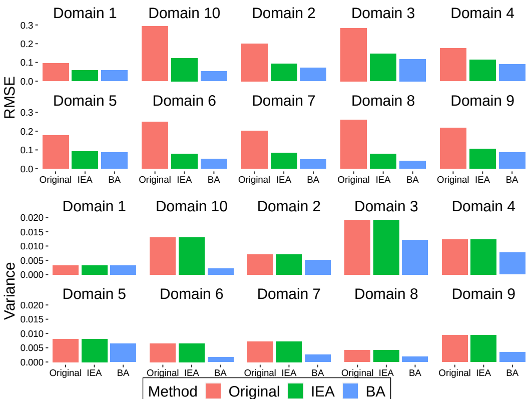

Figure 3: Comparison of the item difficulty estimates among direct applying 2PL-IRT model fitting, item equating adjustment (IEA) and bootstrapping adjustment (BA).

5. CONCLUSION

Nonignorable missingness impacts applying psychometric models to the data collected in VLE. To reduce the impacts of nonignorable missingness, this research explored a statewide-used VLE data to test the hypothesis that the missing values were non-ignorable missingness and related to the factors that 2PL-IRT model measures. The data exploration showed that the non-ignorable missingness would impact the parameter estimation of 2PL-IRT without pre data analysis. To adjust the biased item difficulty parameter estimates caused by the non-ignorable missingness, a semi-supervised learning framework was designed. In the framework, the idea of semi-supervised learning was first time used in IRT area to improve the robustness of latent trait estimation. To convert the observed response pattern to the continuous latent trait and approximate some continuous functions which were hard to specify mathematically, deep learning techniques were also introduced. The combination of semi-supervised deep learning and IRT model improved both accuracy and robustness of the parameter estimation for IRT on noisy data with weak constraint. The experimental results showed that the proposed framework adjust the biases contained in both students' ability estimation and item parameter estimation for 2PL-IRT model.

6. REFERENCES

[1] M. Bienkowski, M. Feng, B. Means, et al. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. *US Department of Education, Office of Educational Technology*, 1:1–57, 2012.

[2] Y. Cui, Q. Guo, and M. Cutumisu. A neural network approach to estimate student skill mastery in cognitive diagnostic assessments. 2017.

[3] S. E. Embretson and S. P. Reise. *Item response theory*. Psychology Press, 2013.

[4] R. Holman and C. A. Glas. Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, 58(1):1–17, 2005.

[5] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[6] F. M. Lord. Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, 48(3):477–482, 1983.

[7] F. Martínez-Plumed, R. B. Prudêncio, A. Martínez-Usó, and J. Hernández-Orallo. Making sense of item response theory in machine learning. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pages 1140–1148. IOS Press, 2016.

[8] B. Means and K. Anderson. Expanding evidence approaches for learning in a digital world. *Office of Educational Technology, US Department of Education*, 2013.

[9] J. Y. Park, T. Dougherty, H. Fritz, and Z. Nagy. Lightlearn: An adaptive and occupant centered controller for lighting based on reinforcement learning. *Building and Environment*, 147:397–414, 2019.

[10] S. Pohl, L. Gräfe, and N. Rose. 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, 2014.

[11] N. Rose, M. Von Davier, and X. Xu. Modeling nonignorable missing data with item response theory (irt). *ETS Research Report Series*, 2010(1):i–53, 2010.

[12] D. B. Rubin. 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, 1987.

[13] K. Taghipour and H. T. Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891, 2016.

[14] M. von Davier. Automated item generation with recurrent neural networks. *psychometrika*, 83(4):847–857, 2018.

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

[16] K. Xue, V. Yaneva, and C. Runyon. On the utility of using transfer learning to predict item characteristics. In *2020 Annual Meeting of the National Council on Measurement in Education (NCME)*, 2020.

[17] K. Xue, V. Yaneva, C. Runyon, and P. Baldwin. Predicting the difficulty and response time of multiple choice questions using transfer learning. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2020.