electronics

Article

Deep Item Response Theory as a Novel Test Theory Based on Deep Learning

Emiko Tsutsumi *, Ryo Kinoshita and Maomi Ueno

Department of Information and Network Engineering, The University of Electro-Communications, Tokyo 182-8585, Japan; Kinoshita@ai.lab.uec.ac.jp (R.K.); ueno@ai.is.uec.ac.jp (M.U.)
* Correspondence: tsutsumi@ai.lab.uec.ac.jp

Abstract: Item Response Theory (IRT) evaluates, on the same scale, examinees who take different tests. It requires the linkage of examinees' ability scores as estimated from different tests. However, the IRT linkage techniques assume independently random sampling of examinees' abilities from a standard normal distribution. Because of this assumption, the linkage not only requires much labor to design, but it also has no guarantee of optimality. To resolve that shortcoming, this study proposes a novel IRT based on deep learning, Deep-IRT, which requires no assumption of randomly sampled examinees' abilities from a distribution. Experiment results demonstrate that Deep-IRT estimates examinees' abilities more accurately than the traditional IRT does. Moreover, Deep-IRT can express actual examinees' ability distributions flexibly, not merely following the standard normal distribution assumed for traditional IRT. Furthermore, the results show that Deep-IRT more accurately predicts examinee responses to unknown items from the examinee's own past response histories than IRT does.

Keywords: deep learning; e-testing; test theory; item response theory

check for updates

Citation: Tsutsumi, E.; Kinoshita, R.; Ueno, M. Deep Item Response Theory as a Novel Test Theory Based on Deep Learning. *Electronics* **2021**, *10*, 1020. https://doi.org/10.3390/electronics10091020

Academic Editor: George A. Tsihrintzis

Received: 22 March 2021
Accepted: 23 April 2021
Published: 25 April 2021

1. Introduction

As a rapidly growing area of e-assessment, E-testing involves the delivery of examinations and assessments on screen, using either local systems or web-based systems. In general, e-testing provides automatic assemblies of uniform test forms, for which each form comprises a different set of items but which still has equivalent measurement accuracy [1–10]. Uniform test forms are assembled for which all forms have equivalent qualities for equal evaluation of examinees who have taken different test forms. Examinees' test scores should be guaranteed to become equivalent, even if different examinees with the same ability take different tests. However because it is difficult to develop perfectly uniform test forms, the calibration process is fundamentally important when multiple test forms are used. To resolve this difficulty, IRT has been used as a calibration method. Reports of the literature describe that Item Response Theory (IRT) offers the following benefits [11,12]:

- One can estimate examinees' abilities while minimizing the effects of heterogeneous or aberrant items that have low estimation accuracy.
- IRT produces examinee ability estimates on a single scale, even for results obtained from different tests.
- IRT predicts an individual examinee's correct response probability to an item from the examinee's past response histories.

Evaluating abilities of numerous examinees on a single scale requires linkage of examinees' abilities estimated from different tests [12–15]. However, linkage techniques of IRT assume random sampling of examinees' abilities from a standard normal distribution. Because of this assumption, the IRT linkage theoretically has no guarantee for its optimality.

Nevertheless, it requires much labor to design [16–19]. In addition, examinees' abilities have no guarantee of being sampled randomly from a standard normal distribution.

To resolve difficulties of linkage, this study proposes a novel Item Response Theory based on deep learning, Deep-IRT, without assuming random sampling of examinees' abilities from a statistical distribution. The proposed method represents a probability for an examinee to answer an item correctly based on the examinee's ability parameter and the item's difficulty parameter. The main contributions of this study are presented below:

- Based on deep learning technology, a novel IRT is proposed. It requires no linkage procedures because it does not assume random sampling of examinees.
- Deep-IRT estimated examinees' abilities with high accuracy when the examinees are not sampled randomly from a single distribution or when there are no common items among the different tests.
- Deep-IRT can express actual examinees' abilities distributions flexibly. It does not follow a standard normal distribution.
- The proposed method provides more reliable and robust ability estimation for actual data than IRT does.

In the study of artificial intelligence, researchers have recently developed deep learning methods that incorporate IRT for knowledge tracing [20–23]. Nevertheless, these methods have not achieved interpretable parameters for examinee ability and item difficulty because each examinee parameter depends on each item. Estimating interpretable parameters is the most important task in the field of test theory. To increase the interpretability of the parameters, the proposed method estimates parameters using two independent networks: an examinee network and an item network. However, generally speaking, independent networks are known to have less prediction accuracy than dependent networks have. Recent studies of deep learning have demonstrated that redundancy of parameters (deep layers of hidden variables) reduces generalization error, contrary to Occam's razor [24–27]. Based on reports of state-of-the-art studies, the proposed method constructs two independent redundant deep networks: an examinee network and an item network. The present study uses the term "deep learning" in the sense of learning neural networks with a deep layer of hidden variables. Therefore, the proposed method is expected to have highly interpretable parameters without impairment of the estimation accuracy.

Simulation experiments demonstrate that the proposed Deep-IRT estimates examinees' abilities more accurately than IRT does when examinees' abilities are not sampled randomly from a single distribution or when no common items exist among the different tests. Experiments conducted with actual data demonstrated that the proposed method provides more reliable and robust ability estimation than IRT does. Furthermore, Deep-IRT more accurately predicted examinee responses to unknown items from the examinee's past response histories than IRT does.

2. Related Works

For knowledge tracing [28–34], the task of tracking the knowledge states of different learners over time, several deep IRT methods that have been developed in the domain of artificial intelligence combine IRT with a deep learning method [20–23,27]. Cheng and Liu [21] proposed deep IRT based Long-short term memory (LSTM) [35] to estimate item discrimination and difficulty parameters by extracting item text information. Yeung [20] and Gan et al. [23] used the dynamic key-value memory network (DKVMN) [21] based on a Memory-Augmented Neural Network and attention mechanisms that trace a learner's knowledge state. Ghosh et al. [22] used attention mechanisms that incorporates a forgetting function of the past learner's response data. Ghosh et al. used a Rasch model [13,36] incorporating the learner's ability parameters and the item's difficulty parameter.

These deep knowledge tracing methods have not achieved interpretable parameters for learner ability and item difficulty, which are extremely important in the field of test theory. In addition, these earlier deep knowledge tracing methods estimate time-series changes of an examinee's abilities to capture the examinee's growth for knowledge tracing.

However, the examinees' ability change is not considered in the field of test theory because the purpose of testing is estimating an examinee's current ability.

Consequently, earlier deep knowledge tracing methods [20–23,27] emphasized not a test theory but a knowledge tracing task. By contrast, this study proposes an IRT model based on deep learning as a novel test theory. Herein, we designate the proposed IRT as "Deep-IRT": a novel test theory.

3. Item Response Theory

This section briefly introduces IRT and a two-parameter logistic model (2PLM), which is an extremely popular IRT model as a test theory. For the two-parameter logistic model, u_{ij} denotes the response of examinee i to item j $(1, \dots, n)$ as

$$u_{ij} = \begin{cases} 1 & (examinee\ i\ answers\ correctly\ to\ item\ j) \\ 0 & (otherwise) \end{cases}$$

In the two parameter logistic model, the probability of a correct answer given to item j by examinee i with ability parameter $\theta_i \in (-\infty, \infty)$ is assumed as

$$\begin{aligned} P_j(\theta_i) &= P(u_{ij} = 1 \mid \theta_i) \\ &= \frac{1}{1 + exp(-1.7a_j(\theta_i - b_j))}, \end{aligned} \tag{1}$$

where $a_j \in (0, \infty)$ is the j-th item's discrimination parameter expressing the discriminatory power for examinee's abilities, and where $b_j \in (-\infty, \infty)$ is the j-th item's difficulty parameter expressing the degree of difficulty. From Bayes' theorem, the posterior distribution of an ability parameter $g(\theta|\boldsymbol{u})$ is given as

$$g(\theta|\boldsymbol{u}) = \frac{L(\theta|\boldsymbol{u})f(\theta)}{h(\boldsymbol{u})}, \tag{2}$$

where $h(\boldsymbol{u})$ is a marginal distribution:

$$h(\boldsymbol{u}) = \int_{-\infty}^{\infty} L(\theta|\boldsymbol{u})f(\theta)d\theta. \tag{3}$$

The parameters are estimated using the expected a priori (EAP) method, which is known to maximize the prediction accuracy theoretically as

$$\hat{\theta} = \int_{-\infty}^{\infty} \theta g(\theta|\boldsymbol{u})d\theta. \tag{4}$$

Because calculating the parameters analytically is difficult, numerical calculation methods such as Markov Chain Monte Carlo methods (MCMC) are generally used.

Here, prior distribution $f(\theta)$ indicates the examinees' ability distribution. The examinees' abilities are assumed to be sampled randomly from $f(\theta)$. Therefore, comparing the examinees' abilities as estimated from different tests requires a linkage that scales those abilities on the same scale using common examinees or items among the tests.

Evaluating examinees' abilities on the same scale requires linkage of examinees' abilities as estimated from different tests [12–15]. Many researchers have developed IRT linkage and calibration methods. Linkage and calibration methods for IRT are divisible into separate calibrations, calibrations with fixed common item parameters, and concurrent calibrations as follows [37–40]:

- Common-item non-equivalent group linkage: transforming scales of parameters into common scales using common items such as substituting the means and deviations of the item parameter estimates of common items [41–49].

- Concurrent calibration: Item parameters for different tests are estimated together using common items [37,50].
- Fixed common item parameters: fixing the common item parameters and calibrating only the pretest items so that the item parameter estimates of the pretest are of the same scale as the common item parameters [51,52].

Even though linkage requires much labor to design, no linkage method can fully represent the joint probability distribution. Particularly when examinees are not sampled randomly from a certain statistical distribution, the linkage accuracy is greatly decreased [16–19]. In addition, examinees' abilities might not be sampled randomly from the standard normal distribution.

4. Deep-IRT

To resolve the difficulties described above, this study proposes a novel Item Response Theory based on Deep Learning: Deep-IRT. To increase the interpretability of the parameters, Deep-IRT estimates parameters using two independent networks: an examinee network and an item network. However, in general, independent networks are known to have less prediction accuracy than dependent networks have. Recent studies of deep learning have demonstrated that redundancy of parameters (deep layers of hidden variables) reduces generalization error, contrary to Occam's razor [24–27]. Based on state-of-the-art reports, Deep-IRT constructs two independent redundant deep networks: an examinee network and an item network. Deep-IRT is expected to have highly interpretable parameters without impairment of the estimation accuracy.

4.1. Method

This subsection presents an explanation of the Deep-IRT method. This method uses two independent neural networks: Examinee Layer and Item Layer. Using outputs of both networks, a probability for an examinee to answer an item correctly is calculated. Figure 1 presents a brief illustration.

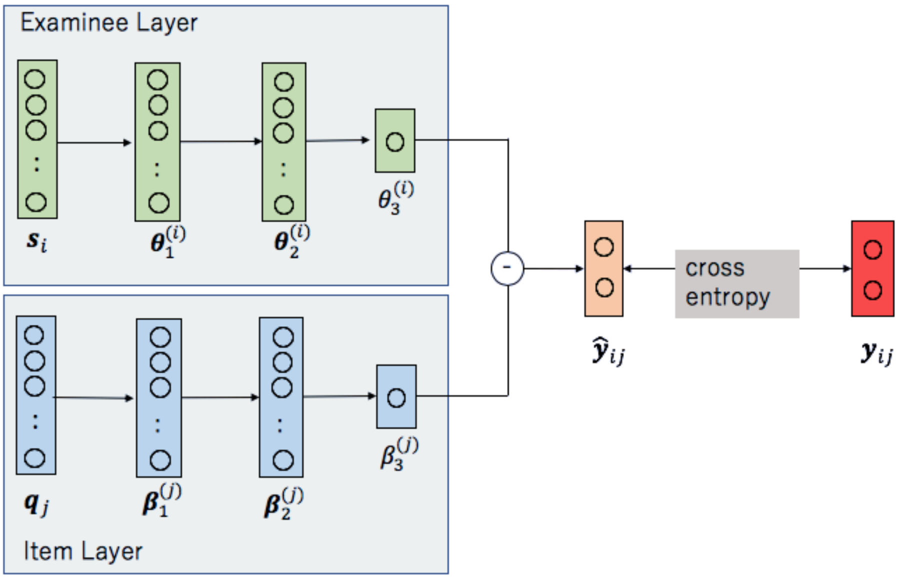


Figure 1. Outline of Deep-IRT.

To express the i-th examinee, the encode of Examinee Layer is a one-hot vector $s_i \in \{0,1\}^I$, where I represents the number of examinees. The i-th element is 1; the other elements are 0s. The Examinee Layer comprises three layers as described below:

$$\theta_1^{(i)} = tanh\left(W^{(\theta_1)} s_i + \tau^{(\theta_1)}\right). \tag{5}$$

$$\boldsymbol{\theta}_2^{(i)} = tanh\left(\boldsymbol{W}^{(\theta_2)}\boldsymbol{\theta}_1^{(i)} + \boldsymbol{\tau}^{(\theta_2)}\right). \tag{6}$$

$$\theta_3^{(i)} = \boldsymbol{W}^{(\theta_3)}\boldsymbol{\theta}_2^{(i)} + \tau^{(\theta_3)}. \tag{7}$$

Here, we use the hyperbolic tangent as an activation function:

$$tanh(x) = \frac{exp(x) - exp(-x)}{exp(x) + exp(-x)}. \tag{8}$$

In addition, $\boldsymbol{W}^{(\theta_1)}$ and $\boldsymbol{W}^{(\theta_2)}$ are the weight matrices given as

$$\boldsymbol{W}^{(\theta_1)} = \begin{pmatrix} w_{11}^{(\theta_1)} & w_{12}^{(\theta_1)} & \dots & w_{1I}^{(\theta_1)} \\ w_{21}^{(\theta_1)} & w_{22}^{(\theta_1)} & \dots & w_{2I}^{(\theta_1)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{\theta}_1|1}^{(\theta_1)} & w_{|\boldsymbol{\theta}_1|2}^{(\theta_1)} & \dots & w_{|\boldsymbol{\theta}_1|I}^{(\theta_1)} \end{pmatrix}$$

$$\boldsymbol{W}^{(\theta_2)} = \begin{pmatrix} w_{11}^{(\theta_2)} & w_{12}^{(\theta_2)} & \dots & w_{1|\boldsymbol{\theta}_1|}^{(\theta_2)} \\ w_{21}^{(\theta_2)} & w_{22}^{(\theta_2)} & \dots & w_{2|\boldsymbol{\theta}_1|}^{(\theta_2)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{\theta}_2|1}^{(\theta_2)} & w_{|\boldsymbol{\theta}_2|2}^{(\theta_2)} & \dots & w_{|\boldsymbol{\theta}_2||\boldsymbol{\theta}_1|}^{(\theta_2)}. \end{pmatrix}$$

Therein, $\boldsymbol{W}^{(\theta_3)}$ is the weight vector given as

$$\boldsymbol{W}^{(\theta_3)} = \left(\ w_1^{(\theta_3)},\ \ w_2^{(\theta_3)},\ \ \dots,\ \ w_{|\boldsymbol{\theta}_2|}^{(\theta_3)}.\ \right)$$

In addition, $\boldsymbol{\tau}^{(\theta_1)} = \left(\tau_1^{(\theta_1)}, \tau_2^{(\theta_1)}, ..., \tau_{|\boldsymbol{\theta}_1|}^{(\theta_1)}\right)$ and $\boldsymbol{\tau}^{(\theta_2)} = \left(\tau_1^{(\theta_2)}, \tau_2^{(\theta_2)}, ..., \tau_{|\boldsymbol{\theta}_2|}^{(\theta_2)}\right)$ are the bias parameters vectors; $\tau^{(\theta_3)}$ is the bias parameter. In this study, we consider the last layer $\theta_3^{(i)}$ as the ith examinee's ability parameter. An overview of the calculation in terms of the Examinee Layer is presented in Figure 2. Weight matrix W represents an estimate of the relation between an examinee's ability and all other examinees' abilities. Therefore, Deep-IRT does not require assumption of random sampling examinees' abilities from a statistical distribution because it estimates an examinees' ability by adjusting the other examinees' ability estimates.

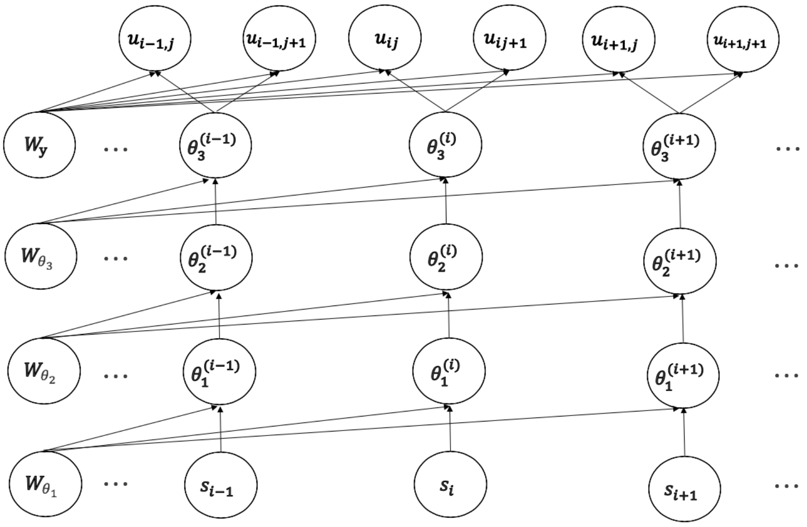


Figure 2. Examinee layer structure.

Similarly, to express the j-th item, the encoding of the Item Layer is a one-hot vector $\boldsymbol{q}_i \in \{0,1\}^J$, where J stands for the number of items. The j-th element is 1; the other elements are 0s. The Item Layer consists of three layers as follows:

$$\boldsymbol{\beta}_1^{(j)} = tanh\Big(\boldsymbol{W}^{(\beta_1)}\boldsymbol{q}_j + \boldsymbol{\tau}^{(\beta_1)}\Big). \tag{9}$$

$$\boldsymbol{\beta}_2^{(j)} = tanh\Big(\boldsymbol{W}^{(\beta_2)}\boldsymbol{\beta}_1^{(j)} + \boldsymbol{\tau}^{(\beta_2)}\Big). \tag{10}$$

$$\beta_3^{(j)} = \boldsymbol{W}^{(\beta_3)}\boldsymbol{\beta}_2^{(j)} + \tau^{(\beta_3)}. \tag{11}$$

In addition, $\boldsymbol{W}^{(\beta_1)}$ and $\boldsymbol{W}^{(\beta_2)}$ are the weight matrices given as presented below:

$$\boldsymbol{W}^{(\beta_1)} = \begin{pmatrix} w_{11}^{(\beta_1)} & w_{12}^{(\beta_1)} & \dots & w_{1J}^{(\beta_1)} \\ w_{21}^{(\beta_1)} & w_{22}^{(\beta_1)} & \dots & w_{2J}^{(\beta_1)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{\beta}_1|1}^{(\beta_1)} & w_{|\boldsymbol{\beta}_1|2}^{(\beta_1)} & \dots & w_{|\boldsymbol{\beta}_1|J}^{(\beta_1)} \end{pmatrix}$$

$$\boldsymbol{W}^{(\beta_2)} = \begin{pmatrix} w_{11}^{(\beta_2)} & w_{12}^{(\beta_2)} & \dots & w_{1|\boldsymbol{\beta}_1|}^{(\beta_2)} \\ w_{21}^{(\beta_2)} & w_{22}^{(\beta_2)} & \dots & w_{2|\boldsymbol{\beta}_1|}^{(\beta_2)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{\beta}_2|1}^{(\beta_2)} & w_{|\boldsymbol{\beta}_2|2}^{(\beta_2)} & \dots & w_{|\boldsymbol{\beta}_2||\boldsymbol{\beta}_1|}^{(\beta_2)} \end{pmatrix}$$

Here, $\boldsymbol{W}^{(\beta_3)}$ is the weight vector given as shown below:

$$\boldsymbol{W}^{(\beta_3)} = \begin{pmatrix} w_1^{(\beta_3)}, & w_2^{(\beta_3)}, & \dots, & w_{|\boldsymbol{\beta}_2|}^{(\beta_3)} \end{pmatrix}$$

Additionally, $\boldsymbol{\tau}^{(\beta_1)} = \Big(\tau_1^{(\beta_1)}, \tau_2^{(\beta_1)}, \dots, \tau_{|\boldsymbol{\beta}_1|}^{(\beta_1)}\Big)$ and $\boldsymbol{\tau}^{(\beta_2)} = \Big(\tau_1^{(\beta_2)}, \tau_2^{(\beta_2)}, \dots, \tau_{|\boldsymbol{\beta}_2|}^{(\beta_2)}\Big)$ are the bias parameters' vectors. $\tau^{(\beta_3)}$ is the bias parameter. For this study, we consider the last layer $\beta_3^{(j)}$ as the j th item's difficulty parameter. Similarly to the sampling of examinees, this method does not assume random sampling of item difficulty parameters from a statistical distribution.

Then, Deep-IRT represents an examinee's correct response probability to an item using the difference between the examinee's ability parameter and the item difficulty parameter. Specifically, examinee i's correct response probability to j's item is described using a hidden layer $\boldsymbol{h}^{(i,j)} = (h_0^{(i,j)}, h_1^{(i,j)})$ as

$$\boldsymbol{h}^{(i,j)} = (\boldsymbol{W}^{(y)})^T(\theta_3^{(i)} - \beta_3^{(j)}) + \boldsymbol{\tau}^{(y)}. \tag{12}$$

$$\begin{aligned} \hat{y}_{i,j} &= softmax(\boldsymbol{h}^{(i,j)}) \\ &= \frac{exp(h_1^{(i,j)})}{exp(h_0^{(i,j)}) + exp(h_1^{(i,j)})}. \end{aligned} \tag{13}$$

Here, $\boldsymbol{W}^{(y)} = (w_1^{(y)}, w_2^{(y)})$ and $\boldsymbol{\tau}^{(y)} = (\tau_1^{(y)}, \tau_2^{(y)})$ are the weight vector and bias' parameters vector.

Deep-IRT does not assume random sampling of examinees' abilities and item difficulties from any statistical distribution. Instead, it uses a deep learning method to estimate the relation between an examinees' ability and all other examinees' abilities by maximizing the prediction accuracy of examinees' responses. The unique feature of this method is to estimate an examinee's ability by adjusting the other examinees' ability estimates. Because of this property, this method requires no linkage procedure.

4.2. Learning Parameters

In general, deep learning methods learn their parameters using the back-propagation algorithm by minimizing a loss function. The loss function of the proposed Deep-IRT employs cross-entropy, which reflects classification errors. It is calculated from the predicted responses $\hat{y}_{i,j}$ and the true responses $u_{i,j}$ as

$$\ell(u_{i,j}, \hat{y}_{i,j}) = -u_{i,j} \log \hat{y}_{i,j} - (1 - u_{i,j}) \log(1 - \hat{y}_{i,j}). \tag{14}$$

Like other machine learning techniques, deep learning methods are biased to data they have encountered before. Therefore, the generalization capacity of the methods depends on the training data, which leads to sub-optimal performance. Consequently, Deep-IRT cannot predict responses of examinees or items accurately with an extremely small number of (in)correct answers. To overcome this shortcoming, cost-sensitive learning, which weights minority data over majority, has been used widely [53]. Therefore, we add the loss function based on a cost-sensitive approach as

$$\begin{aligned} Loss_{class} = & \sum_i \sum_j \ell(u_{i,j}, \hat{y}_{i,j}) \\ & + \gamma_1 \sum_{i \in L_e} \sum_{j \in (u_{i,j}=1)} \ell(u_{i,j}, \hat{y}_{i,j}) \\ & + \gamma_2 \sum_{i \in H_e} \sum_{j \in (u_{i,j}=0)} \ell(u_{i,j}, \hat{y}_{i,j}) \\ & + \gamma_3 \sum_{j \in L_i} \sum_{i \in (u_{i,j}=1)} \ell(u_{i,j}, \hat{y}_{i,j}) \\ & + \gamma_4 \sum_{j \in H_i} \sum_{i \in (u_{i,j}=0)} \ell(u_{i,j}, \hat{y}_{i,j}), \end{aligned} \tag{15}$$

where L_e stands for a group of examinees whose correct answer rates are less than α_{L_e}, H_e denotes a group of examinees whose correct answer rates are more than α_{H_e}, L_i signifies a group of items of which correct answer rates are less than α_{L_i}, and H_i represents a group of items with correct answer rates that are more than α_{H_i}. Here, $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ and $\alpha_{L_e}, \alpha_{H_e}, \alpha_{L_i}, \alpha_{H_i}$ are tuning parameters.

All of the parameters are learned simultaneously using a popular optimization algorithm: adaptive moment estimation [54].

5. Simulation Experiments

This section presents evaluation of the performances of Deep-IRT using simulation data according to earlier IRT studies of the linkage or the multi-population [55,56].

5.1. Experiment Settings

We implemented Deep-IRT using Chainer (https://chainer.org/ (accessed on 23 April 2021)), a popular framework for neural networks. The values of tuning parameters are presented in Table 1.

For implementation of IRT, we employ 2PLM and estimate the parameters using EAP estimation with the MCMC algorithm. The prior distributions are

$$\theta \sim N(0, 1), \log a \sim N(0, 1), b \sim N(1, 0.4). \tag{16}$$

Table 1. Values of tuning parameters.

Parameter	Value	Parameter	Value
$\lvert\boldsymbol{\theta}_1^{(i)}\rvert$	50	γ_1	0.1
$\lvert\boldsymbol{\theta}_2^{(i)}\rvert$	50	γ_2	0.1
$\lvert\boldsymbol{\beta}_1^{(j)}\rvert$	50	γ_3	0.1
$\lvert\boldsymbol{\beta}_2^{(j)}\rvert$	50	γ_4	0.1
Epoch	300	α_{L_e}	0.2
-	-	α_{H_e}	0.8
-	-	α_{L_i}	0.2
-	-	α_{H_i}	0.8

5.1.1. Estimation Accuracy

For this experiment, we evaluate root mean square error (RMSE), Pearson's correlation coefficient, and the Kendall rank correlation coefficient between the estimated abilities and the true values. For calculation of RMSE, the estimated abilities of Deep-IRT are standardized.

5.2. Estimation Accuracy for Randomly Sampled Examinee Data

To underscore the effectiveness of Deep-IRT for data of examinees' abilities that are not randomly sampled, this subsection presents evaluation of the estimation accuracy with changing examinee assignments for different tests. The procedures of this experiment are explained hereinafter.

This experiment generates 10 test data that have no common examinees. In addition, the k-th test ($k = 1, \ldots, 10$) has common items only among the $k-1$-th test and the $k+1$-th test.

The true parameters were generated randomly:

$$\theta \sim N(0,1), \log a \sim N(0,1), b \sim N(1,0.4). \tag{17}$$

Here, the simulation data were generated based on 2PLM in the following two ways. The first way is that examinees are assigned randomly to each test from Equation (17). The other way is that examinees are assigned systematically to each test as described below.

1. Examinees are sampled randomly from Equation (17).
2. The examinees are sorted in order of their ascending ability. Furthermore, the examinees are divided equally into groups of 10 examinees in order of their respective abilities.
3. The k th examinee group is assigned to the k-th test.

Table 2 demonstrates the average of estimation accuracies for each condition. Results of the random assignment condition show that IRT outperforms Deep-IRT. The reason is that the condition is an ideal situation for IRT because the data are generated randomly from the IRT model. However, for a small number of examinees or items, the differences between IRT and Deep-IRT become smaller.

In contrast, the results obtained for the systematically assignment condition show Deep-IRT without assuming randomly sampling examinees outperforms IRT with that assumption. Furthermore, Deep-IRT suppresses the decline of accuracy in cases without common items among different tests. These results are expected to be beneficial for applying Deep-IRT with actual data.

Table 2. Parameter estimation accuracies.

Assignment	No. Items of Each Test	No. Common Items (Total No. Items)	No. Examinees for Each Test (Total No. Examinees)	Method	RMSE	Pearson	Kendall
random	10	5 (55)	50 (500)	Deep-IRT	0.469	0.890	0.748
				IRT	**0.420**	**0.912**	**0.781**
			100 (1000)	Deep-IRT	0.447	0.900	0.766
				IRT	**0.438**	**0.904**	**0.770**
			500 (5000)	Deep-IRT	0.434	0.907	0.769
				IRT	**0.432**	0.907	**0.776**
			1000 (10,000)	Deep-IRT	0.424	0.908	**0.771**
				IRT	**0.411**	**0.911**	0.733
		0 (100)	50 (500)	Deep-IRT	0.458	0.896	0.747
				IRT	**0.456**	0.896	**0.751**
			100 (1000)	Deep-IRT	0.455	0.832	0.765
				IRT	**0.440**	**0.903**	**0.767**
			500 (5000)	Deep-IRT	0.433	0.852	0.785
				IRT	**0.423**	**0.861**	**0.789**
			1000 (10,000)	Deep-IRT	0.412	0.910	**0.799**
				IRT	**0.403**	**0.914**	0.794
	30	5 (255)	50 (500)	Deep-IRT	0.328	0.921	0.855
				IRT	**0.301**	**0.941**	**0.865**
			100 (1000)	Deep-IRT	0.319	0.949	0.865
				IRT	**0.292**	**0.957**	**0.870**
			500 (5000)	Deep-IRT	0.339	0.942	0.834
				IRT	**0.290**	**0.958**	**0.873**
			1000 (10,000)	Deep-IRT	0.329	0.947	0.844
				IRT	**0.298**	**0.968**	**0.879**
		0 (300)	50 (500)	Deep-IRT	0.328	0.946	**0.860**
				IRT	**0.308**	**0.952**	0.858
			100 (1000)	Deep-IRT	0.339	0.943	0.851
				IRT	**0.314**	**0.951**	**0.858**
			500 (5000)	Deep-IRT	0.321	0.941	0.853
				IRT	**0.299**	**0.945**	**0.873**
			1000 (10,000)	Deep-IRT	0.302	0.938	0.853
				IRT	**0.281**	**0.948**	**0.881**
	50	5 (455)	50 (500)	Deep-IRT	0.317	0.950	0.882
				IRT	**0.251**	**0.969**	**0.895**
			100 (1000)	Deep-IRT	0.312	0.964	0.891
				IRT	**0.243**	**0.970**	**0.896**
			500 (5000)	Deep-IRT	0.288	0.959	0.894
				IRT	**0.232**	**0.973**	**0.901**
			1000 (10,000)	Deep-IRT	0.278	0.961	0.894
				IRT	**0.234**	**0.973**	**0.901**
		0 (500)	50 (500)	Deep-IRT	0.360	0.935	0.856
				IRT	**0.274**	**0.962**	**0.876**
			100 (1000)	Deep-IRT	0.261	0.966	0.884
				IRT	**0.251**	**0.968**	**0.892**
			500 (5000)	Deep-IRT	0.341	0.942	0.887
				IRT	**0.241**	**0.971**	**0.899**
			1000 (10,000)	Deep-IRT	0.266	0.968	0.889
				IRT	**0.241**	**0.972**	**0.901**

Table 2. *Cont.*

Assignment	No. Items of Each Test	No. Common Items (Total No. Items)	No. Examinees for Each Test (Total No. Examinees)	Method	RMSE	Pearson	Kendall
system	10	5 (55)	50 (500)	Deep-IRT	**0.665**	**0.778**	**0.568**
				IRT	1.111	0.381	0.237
			100 (1000)	Deep-IRT	**0.622**	**0.807**	**0.629**
				IRT	0.779	0.696	0.466
			500 (5000)	Deep-IRT	**0.611**	**0.812**	**0.639**
				IRT	0.792	0.702	0.499
			1000 (10,000)	Deep-IRT	**0.621**	**0.822**	**0.651**
				IRT	0.712	0.702	0.501
		0 (100)	50 (500)	Deep-IRT	**0.997**	**0.502**	**0.267**
				IRT	1.170	0.314	0.184
			100 (1000)	Deep-IRT	**0.721**	**0.740**	**0.561**
				IRT	1.176	0.308	0.197
			500 (5000)	Deep-IRT	**0.701**	**0.761**	**0.591**
				IRT	1.016	0.498	0.277
			1000 (10,000)	Deep-IRT	**0.698**	**0.782**	**0.591**
				IRT	0.808	0.673	0.457
	30	5 (255)	50 (500)	Deep-IRT	**0.561**	**0.835**	**0.696**
				IRT	0.613	0.786	0.622
			100 (1000)	Deep-IRT	**0.501**	**0.875**	**0.716**
				IRT	0.573	0.836	0.672
			500 (5000)	Deep-IRT	**0.499**	**0.878**	**0.722**
				IRT	0.553	0.846	0.679
			1000 (10,000)	Deep-IRT	**0.495**	**0.892**	**0.731**
				IRT	0.534	0.851	0.691
		0 (300)	50 (500)	Deep-IRT	**0.661**	**0.781**	**0.586**
				IRT	0.786	0.691	0.489
			100 (1000)	Deep-IRT	**0.579**	**0.832**	**0.664**
				IRT	0.762	0.709	0.506
			500 (5000)	Deep-IRT	**0.561**	**0.852**	**0.684**
				IRT	0.732	0.705	0.512
			1000 (10,000)	Deep-IRT	**0.539**	**0.850**	**0.644**
				IRT	0.712	0.709	0.506
	50	5 (455)	50 (500)	Deep-IRT	**0.376**	**0.929**	**0.802**
				IRT	0.426	0.909	0.760
			100 (1000)	Deep-IRT	**0.393**	**0.923**	**0.811**
				IRT	0.805	0.750	0.543
			500 (5000)	Deep-IRT	**0.372**	**0.930**	**0.810**
				IRT	1.044	0.454	0.282
			1000 (10,000)	Deep-IRT	**0.392**	**0.914**	**0.798**
				IRT	0.923	0.512	0.342
		0 (500)	50 (500)	Deep-IRT	**0.635**	**0.798**	**0.599**
				IRT	0.782	0.694	0.489
			100 (1000)	Deep-IRT	**0.408**	**0.916**	**0.785**
				IRT	0.612	0.812	0.532
			500 (5000)	Deep-IRT	**0.421**	**0.891**	**0.765**
				IRT	0.598	0.822	0.495
			1000 (10,000)	Deep-IRT	**0.411**	**0.901**	**0.785**
				IRT	0.602	0.829	0.498

5.3. Estimation Accuracy for Multi-Population Data

As described earlier, IRT assumes that examinees' abilities follow a standard normal distribution. Furthermore, it is known that no optimal linkage occurs under the assumption [17]. Additionally, no guarantee exists that examinees' abilities follow a standard normal distribution. When the assumption is violated, ability estimation accuracy of IRT becomes extremely worse, even without the linkage problem. However, because Deep-IRT does not assume random sampling from a statistical distribution, robust ability estimation is expected to be provided even when the IRT presumption is violated. To demonstrate the benefits of the proposed method, this subsection evaluates estimation accuracies of IRT and Deep-IRT when examinees' abilities follow multiple populations.

For this experiment, the abilities of examinees taking different tests are assumed to be sampled from different populations. For this study, we assume two tests including 50 items. The abilities of the tests are sampled randomly from $N_1(\mu_1, \sigma^2)$ and $N_2(\mu_2, \sigma^2)$.

Table 3 shows the average of estimation accuracies with different ability distributions and the number of common items. The standard deviation of each distribution was ascertained so that the total abilities' standard deviation is close to 1.0. Here, Wilcoxon's signed rank test is applied to infer whether the accuracies of IRT and Deep-IRT are significantly different. The results showed that when the difference between μ_1 and μ_2 becomes small, IRT provides significantly high accuracy because the distribution approaches a single normal distribution. By contrast, as the difference between μ_1 and μ_2 becomes large, Deep-IRT estimates examinees' abilities accurately. Therefore, Deep-IRT is robust for estimation of examinees' abilities when they follow different distributions. The results also show that, when there is no common item, Deep-IRT estimates the examinees' abilities more accurately than IRT does. Consequently, Deep-IRT can estimate examinees' abilities accurately without common items.

Table 3. Estimation accuracies for multi-population data.

No. Examinees for Each Test	No. Common Items	μ_1	μ_2	σ^2	IRT	Deep-IRT
500	5	−0.3	0.3	0.7	**0.186** **	0.216
		−0.5	0.5	0.5	**0.184** **	0.232
		−0.7	0.7	0.3	0.210	**0.206**
		−0.9	0.9	0.1	0.207	**0.195** *
	0	−0.3	0.3	0.7	0.358	**0.325** *
		−0.5	0.5	0.5	0.501	**0.324** **
		−0.7	0.7	0.3	0.993	**0.382** **
		−0.9	0.9	0.1	1.027	**0.385** **

** $p < 0.01$, * $p < 0.05$.

Next, we demonstrate that Deep-IRT can accommodate abilities with multiple populations. Specifically, we generate abilities according to multiple populations for data $N_1(-0.7, 0.3)$ and $N_2(0.7, 0.3)$ in Table 3. Figure 3 shows histograms of the true abilities, the estimated abilities using IRT, and the estimated abilities using Deep-IRT. Figure 3 shows that Deep-IRT clearly estimates a bimodal distribution as the ability distribution similar to the true distribution. The result demonstrates that Deep-IRT flexibly expresses actual examinees' abilities distributions that do not follow a standard normal distribution.

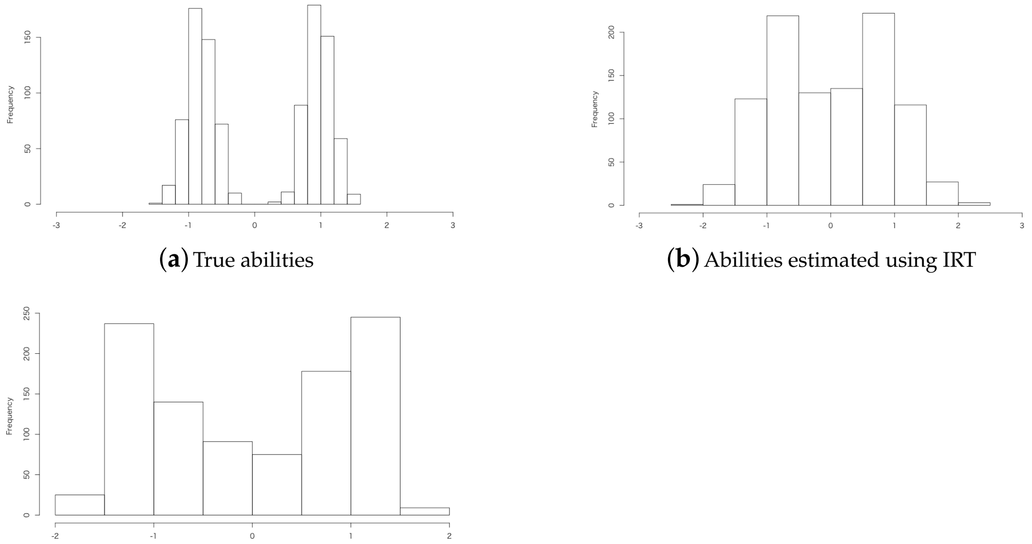
Next, we evaluate the estimated ability distributions of IRT and Deep-IRT using a fitting score to the true distribution as

$$\sum_{k \in \{1,2\}} \sum_{i=1}^{I_k} \log p(\hat{\theta}_{ki} | \mu_k, \sigma), \tag{18}$$

where I_k represents the number of examinees who took the k-th test. In addition, $\hat{\theta}_{ki}$ is the estimated ability of i-th examinee for the k-th test. In addition, $p(\hat{\theta}_{ki}|\mu_k, \sigma)$ is the likelihood of estimated abilities given the true ability distribution as

$$p(\hat{\theta}_{ki}|\mu_k, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\hat{\theta}_{ki} - \mu_k)^2}{2\sigma^2} \right). \tag{19}$$

If the method fits the true distribution, then the estimated distribution approaches the true distribution. The fitting score of IRT is −1633.4. That of Deep-IRT is −1437.1. The latter is higher than the former. Therefore, Deep-IRT expresses the examinees' ability distributions more accurately than IRT does.



(**a**) True abilities

(**b**) Abilities estimated using IRT

(**c**) Abilities estimated using Deep-IRT

Figure 3. Histograms of estimated abilities for multi-population data.

6. Actual Data Experiments

The simulation experiments suggested that Deep-IRT might estimate examinees' abilities with high accuracy for actual data. This section evaluates the effectiveness of Deep-IRT using actual datasets.

6.1. Actual Datasets

For this experiment, we use the following actual datasets. Here, we present "Rate.Sparse" which is the average rate of items that an examinee did not address in the learning process.

1. Information datasets consist of two test data (Information 1, 2) related to information technology. Information 1 has 169 examinees over 50 items. Information 2 has 266 examinees over 50 items. The tests were conducted of the learning management system, "Samurai" developed by [57–59]. Rate.Sparse is 0%.
2. The critical thinking dataset has 1221 undergraduate examinees over 179 items about critical thinking. Rate.Sparse is 87.8%.
3. Program datasets consist of two test data (Program 1, 2) about programming. Program 1 has 93 examinees over 13 items. Rate.Sparse is 0%. Program 2 has 74 examinees over 19 items with 6.8% Rate.Sparse.
4. Practice Exam dataset consists of two test data for high school students. Each test relates to mathematics and physics. Mathematics data have 12,348 examinees over 48 items. Physics data have 9172 examinees over 24 items. The respective values of Rate.Sparse are 16.4% and 12.0%.

5. Assistments dataset is the 2009–2010 dataset of Assistments (https://sites.google.com/site/assistmentsdata/\home/assistment-2009-2010-data (accessed on 23 April 2021).), which is a large dataset that is used widely for knowledge tracing. Here, we removed examinees answering only one item and items answered by fewer than 30 examinees. For that reason, our dataset has 3941 examinees over 2921 items with 84.4% Rate.Sparse.
6. CDM datasets, which are widely used open datasets, are included in the R package CDM [60]. We used two datasets: ECPE and TIMSS. ECPE data include those for 2922 examinees over 28 language-related items. TIMSS data include those for 757 examinees over 23 math items. Rate.Sparse is 0%.
7. Statistics dataset includes those of 26 undergraduate examinees over 25 items about statistics. Rate.Sparse is 33.8%.
8. Information Ethics dataset has 31 undergraduate examinees over 90 items related to information ethics. Rate.Sparse is 46.3%.
9. Engineer Ethics dataset has 85 undergraduate examinees over 69 items related to engineer ethics. Rate.Sparse is 26.4%.
10. Classi datasets consist of three test data for high school examinees: tests relate to physics, chemistry, and biology. The tests were conducted on the web-based system, "Classi (https://classi.jp (accessed on 23 April 2021).)" using a tablet. Datasets have 239, 1139, and 192 examinees, respectively, and 119, 364, and 114 items. The respective values of Rate.Sparse are 92.4%, 96.4%, and 93.5%.

6.2. Reliability of Ability Estimation

This subsection presents evaluation of the reliability of abilities estimation of Deep-IRT. Because the true values of parameters are unknown, we evaluate the reliabilities as follows: (1) Each dataset is divided equally into two sets of data. (2) Parameters of each method are estimated for the divided data from each dataset. (3) The RMSE and correlation between the two sets of the estimated parameters from the two divided datasets are calculated. (4) These procedures are repeated 10 times. The average of the RMSEs and correlations is calculated. Table 4 presents the results. Here, a Wilcoxon signed rank test is applied to infer whether the reliabilities of IRT and Deep-IRT are significantly different.

Table 4. Reliability of ability parameter estimation.

Dataset	Method	RMSE	Pearson	Kendall
Information 1	2PLM	**0.466**	**0.891**	**0.685**
	Deep-IRT	0.514	0.867	0.687
Information 2	2PLM	0.562	0.841	0.668
	Deep-IRT	**0.555**	**0.845**	**0.662**
Critical Thinking	2PLM	1.064	0.464	0.318
	Deep-IRT	**1.025**	**0.474**	**0.327**
Program 1	2PLM	0.890	0.599	0.403
	Deep-IRT	**0.864**	**0.622**	**0.417**
Program 2	2PLM	0.752	0.713	0.468
	Deep-IRT	**0.720**	**0.737**	**0.475**
Practice_Math	2PLM	**0.589**	**0.748**	**0.533**
	Deep-IRT	0.744	0.723	0.514
Practice_Physics	2PLM	**0.884**	**0.609**	**0.424**
	Deep-IRT	0.911	0.585	0.411

Table 4. *Cont.*

Dataset	Method	RMSE	Pearson	Kendall
ASSISTMENTS	2PLM	**0.827**	**0.658**	**0.441**
	Deep-IRT	0.849	0.639	0.478
ECPE	2PLM	0.875	0.615	0.435
	Deep-IRT	**0.874**	**0.618**	**0.440**
TIMSS	2PLM	0.753	0.716	0.525
	Deep-IRT	0.753	0.716	**0.523**
Statistics	2PLM	0.619	0.801	0.398
	Deep-IRT	**0.545**	**0.846**	**0.582**
Information Ethics	2PLM	0.394	0.920	0.643
	Deep-IRT	**0.382**	**0.925**	**0.712**
Engineer Ethics	2PLM	0.544	0.850	0.403
	Deep-IRT	**0.517**	**0.865**	**0.313**
Classi_Physics	2PLM	1.053	0.444	0.299
	Deep-IRT	**0.943**	**0.554**	**0.403**
Classi_Chemistry	2PLM	1.077	0.420	0.297
	Deep-IRT	**0.923**	**0.574**	**0.439**
Classi_Biology	2PLM	1.020	0.475	0.326
	Deep-IRT	**0.748**	**0.717**	**0.531**
Average	2PLM	0.764	0.680	0.451
	Deep-IRT	**0.742**	**0.707**	**0.495** *

* $p < 0.05$.

Table 4 shows that Deep-IRT provides more reliable ability estimates than IRT does. In particular, regarding the average of Kendall rank correlation coefficient, which is known to provide a robust estimate for aberrant values, Deep-IRT outperforms IRT significantly. Results indicate that Deep-IRT can estimate parameters more reliably than IRT does for actual test data. It is surprising that Deep-IRT outperforms IRT for small datasets such as Program 1, Program 2, Statistics, Information Ethics, and Engineer Ethics. This result indicates Deep-IRT as effective even for small datasets. For Practice_Math, Practice_Physics, and ASSISTMENTS, IRT has a higher Kendall rank correlation coefficient than Deep-IRT does because the ability estimation of IRT tends to become stable when the dataset becomes large. IRT has that stability because it is guaranteed to converge asymptotically to the true joint probability distribution.

6.3. Prediction of Responses to Unknown Items

In the field of artificial intelligence in education, the prediction of examinee's responses to unknown items from the examinee's past response history becomes important for adaptive learning systems [20,30,32,61,62]. Reportedly, the prediction accuracy of IRT is the highest for the problem [63]. This subsection presents comparison of the prediction accuracy of Deep-IRT with that of IRT. Specifically, using ten-fold cross validation, the parameters are learned from training data and are used to predict responses in the remaining data. Then, we calculate the accuracy rates for the cross validation experiments. Here, a Wilcoxon signed rank test is applied to infer whether the respective accuracies of IRT and Deep-IRT are significantly different.

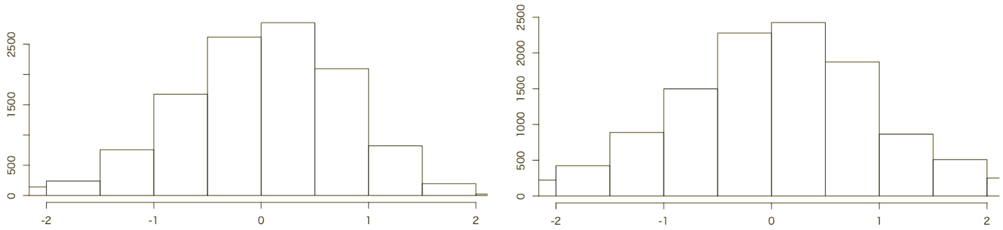
Table 5 shows the results: the average of F1 value of Deep-IRT is significantly higher than that of IRT. Deep-IRT can predict examinees' responses to unknown items more accurately than IRT can. It is noteworthy that Deep-IRT does not always outperform for large data. For ASSISTMENTS and Critical Thinking, IRT provides better performance than Deep-IRT does because ASSISTMENTS and Critical Thinking have high values of Rate.Sparse. Deep-IRT might be weak in dealing with sparse datasets. In contrast, for datasets with low values of Rate.Sparse, Deep-IRT outperforms IRT even for small datasets. Generally speaking, the IRT prediction accuracy increases along with the number of examinees. Therefore, IRT has high prediction accuracies for Practice_Math and Practice_Physics.

Table 5. Prediction accuracies of responses to unknown items.

Data	No. Examinees	No. Items	Rate.Sparse	IRT	Deep-IRT
Information 1	169	50	0%	0.734	**0.737**
Information 2	266	50	0%	0.699	**0.700**
Critical Thinking	1221	179	87.8%	**0.695**	0.689
Program 1	94	13	0%	0.719	**0.729**
Program 2	74	19	6.8%	0.676	**0.685**
Practice_Math	12,348	48	16.4%	**0.783**	0.780
Practice_Physics	9172	24	12.0%	**0.721**	0.710
ASSISTMENTS	3941	2921	84.4%	**0.685**	0.679
ECPE	2922	28	0%	0.719	**0.729**
TIMSS	757	24	0%	0.711	**0.712**
Statistics	26	25	33.8%	0.852	**0.893**
Information Ethics	31	90	46.3%	0.746	**0.803**
Engineer Ethics	85	69	26.4%	0.634	**0.685**
Classi_Physics	239	119	92.4%	0.720	**0.721**
Classi_Chemistry	1139	364	96.4%	0.710	**0.711**
Classi_Biology	192	114	93.5%	0.722	**0.725**
Average				0.719	**0.728** *

* $p < 0.05$.

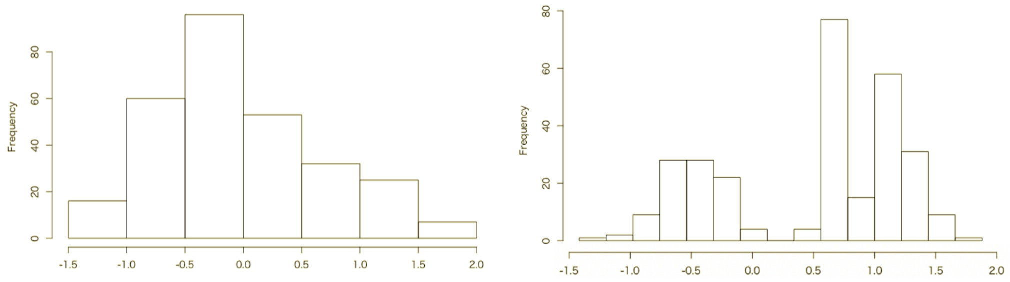
Furthermore, Figure 4 depicts histograms of abilities estimated from Practice_Math, where the prediction accuracy of IRT is higher than that of Deep-IRT. Figure 5 depicts histograms of abilities estimated from Classi_Biology data, where the prediction accuracy of Deep-IRT is higher than that of IRT. Figure 4 shows estimates conducted using both methods for the ability distribution similar to the standard normal distribution. In contrast, Figure 5 shows that Deep-IRT expresses a multi modal distribution, although IRT estimates a unimodal distribution. Deep-IRT can predict responses to unknown items because it can flexibly express distributions of various abilities.



(**a**) Abilities estimated using IRT

(**b**) Abilities estimated using Deep-IRT

Figure 4. Histograms of abilities estimated using IRT and Deep-IRT for Practice_Math data.



(**a**) Abilities estimated using IRT (**b**) Abilities estimated using Deep-IRT

Figure 5. Histograms of abilities estimated using IRT and Deep-IRT for Classi_Biology data.

7. Conclusions

This study examines a novel test theory based on deep learning: Deep-IRT. To increase the interpretability of the parameters, Deep-IRT estimates parameters using two independent networks: an examinee network and an item network. However, generally speaking, independent networks are known to have less prediction accuracy than dependent networks have. Recent studies of deep learning have indicated that redundancy of parameters reduces generalization error, contrary to Occam's razor [24–27]. Based on reports of state-of-the-art research, Deep-IRT was constructed to have two independent redundant deep networks. Therefore, Deep-IRT has high interpretable parameters without impairment of the estimation accuracy. The main contributions of Deep-IRT are presented below:

(1) Deep-IRT does not assume random sampling of examinees' abilities from a statistical distribution because the weight matrix of the ability parameters estimates the relation between an examinee's ability and all other examinees' abilities.

(2) Deep-IRT estimates examinees' abilities with high accuracy when the examinees are not sampled randomly from a single distribution or when no common items exist among the different tests.

(3) Deep-IRT flexibly expresses actual examinees' ability distributions that do not follow a standard normal distribution.

Experiments conducted using actual data demonstrated that Deep-IRT provided more reliable and robust ability estimation than IRT did. Furthermore, Deep-IRT more accurately predicted examinee responses to unknown items from the examinee's past response histories than IRT did. Results showed that Deep-IRT is effective even for small datasets. However, the results also suggest that Deep-IRT might be weak in dealing with sparse data. To estimate an examinee's ability for sparse data robustly, one must improve the estimation methods. One potential means of doing so is optimizing the number of hidden layers of each neural network.

Furthermore, as another subject of future work, we expect to incorporate Deep-IRT with (CAT) [64,65] to improve the examinee's ability estimation accuracy in an actual environment.

Author Contributions: Conceptualization, methodology, E.T., R.K., and M.U.; validation, R.K.; writing—original draft preparation, E.T. and R.K.; writing—review and editing, M.U.; funding acquisition, M.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JSPS KAKENHI Grant Numbers JP19H05663 and JP19K21751.

Conflicts of Interest: The authors declare that they have no conflict of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

1. Songmuang, P.; Ueno, M. Bees Algorithm for Construction of Multiple Test Forms in E-Testing. *IEEE Trans. Learn. Technol.* **2011**, *4*, 209–221. [CrossRef]
2. Ishii, T.; Songmuang, P.; Ueno, M. Maximum Clique Algorithm for Uniform Test Forms Assembly. In Proceedings of the 16th International Conference on Artificial Intelligence in Education, Memphis, TN, USA, 9–13 July 2013; Volume 7926, pp. 451–462._46. [CrossRef]
3. Ishii, T.; Songmuang, P.; Ueno, M. Maximum Clique Algorithm and Its Approximation for Uniform Test Form Assembly. *IEEE Trans. Learn. Technol.* **2014**, *7*, 83–95. [CrossRef]
4. Ishii, T.; Ueno, M. Clique Algorithm to Minimize Item Exposure for Uniform Test Forms Assembly. In Proceedings of the International Conference on Artificial Intelligence in Education, Madrid, Spain, 22–26 June 2015; pp. 638–641. [CrossRef]
5. Ishii, T.; Ueno, M. Algorithm for Uniform Test Assembly Using a Maximum Clique Problem and Integer Programming. In Proceedings of the Artificial Intelligence in Education, Wuhan, China, 28 June–1 July 2017; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; pp. 102–112. [CrossRef]
6. Lin, Y.; Jiang, Y.S.; Gong, Y.J.; Zhan, Z.H.; Zhang, J. A Discrete Multiobjective Particle Swarm Optimizer for Automated Assembly of Parallel Cognitive Diagnosis Tests. *IEEE Trans. Cybern.* **2018**, 1–14. [CrossRef] [PubMed]
7. Vie, J.J.; Popineau, F.; Bruillard, E.; Bourda, Y. Automated Test Assembly for Handling Learner Cold-Start in Large-Scale Assessments. *Int. J. Artif. Intell. Educ.* **2018**, *28*. [CrossRef]
8. RodrÃguez-Cuadrado, J.; Delgado-GÃ³mez, D.; Laria, J.; Rodriguez-Cuadrado, S. Merged Tree-CAT: A fast method for building precise Computerized Adaptive Tests based on Decision Trees. *Expert Syst. Appl.* **2019**, *143*, 113066. [CrossRef]
9. Linden, W.; Jiang, B. A Shadow-Test Approach to Adaptive Item Calibration. *Psychometrika* **2020**, *85*. [CrossRef]
10. Ren, H.; Choi, S.; Linden, W. Bayesian adaptive testing with polytomous items. *Behaviormetrika* **2020**, *47*. [CrossRef]
11. Lord, F.; Novick, M. *Statistical Theories of Mental Test Scores*; Addison-Wesley: Boston, MA, USA, 1968; p. xiii, 274p.
12. Van der Linden, W.J. *Handbook of Item Response Theory, Volume Three: Applications*; Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences; Chapman and Hall/CRC: Boca Raton, FL, USA, 2016.
13. Lord, F. *Applications of Item Response Theory to Practical Testing Problems*; L. Erlbaum Associates: Hillsdale, NJ, USA, 1980.
14. Van der Linden, W.J. *Handbook of Item Response Theory, Volume Two: Statistical Tools*; Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences; Chapman and Hall/CRC: Boca Raton, FL, USA, 2016.
15. Joo, S.H.; Lee, P.; Stark, S. Evaluating Anchor-Item Designs for Concurrent Calibration With the GGUM. *Appl. Psychol. Meas.* **2017**, *41*, 83–96. [CrossRef]
16. Ogasawara, H. Standard Errors of Item Response Theory Equating/Linking by Response Function Methods. *Appl. Psychol. Meas.* **2001**, *25*, 53–67. [CrossRef]
17. van der Linden, W.; Barrett, M.D. Linking Item Response Model Parameters. *Psychometrika* **2016**, *81*, 650–673. [CrossRef]
18. Andersson, B. Asymptotic Variance of Linking Coefficient Estimators for Polytomous IRT Models. *Appl. Psychol. Meas.* **2018**, *42*, 192–205. [CrossRef]
19. Barrett, M.D.; van der Linden, W.J. Estimating Linking Functions for Response Model Parameters. *J. Educ. Behav. Stat.* **2019**, *44*, 180–209. [CrossRef]
20. Yeung, C. Deep-IRT: Make Deep Learning Based Knowledge Tracing Explainable Using Item Response Theory. In Proceedings of the 12th International Conference on Educational Data Mining, EDM, Montreal, QC, Canada, 2–5 July 2019.
21. Cheng, S.; Liu, Q. Enhancing Item Response Theory for Cognitive Diagnosis. *CoRR* **2019**, abs/1905.10957. Available online: http://xxx.lanl.gov/abs/1905.10957 (accessed on 23 April 2021).
22. Ghosh, A.; Heffernan, N.; Lan, A.S. Context-Aware Attentive Knowledge Tracing. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, CA, USA, 23–27 August 2020.
23. Gan, W.; Sun, Y.; Sun, Y. Knowledge Interaction Enhanced Knowledge Tracing for Learner Performance Prediction. In Proceedings of the 2020 Seventh International Conference on Behavioural and Social Computing (BESC), Bournemouth, UK, 5–7 November 2020; pp. 1–6. [CrossRef]
24. He, H.; Huang, G.; Yuan, Y. Asymmetric Valleys: Beyond Sharp and Flat Local Minima. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: New York City, NY, USA, 2019; pp. 2553–2564.
25. Morcos, A.; Yu, H.; Paganini, M.; Tian, Y. One ticket to win them all: Generalizing lottery ticket initializations across datasets and optimizers. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: New York City, NY, USA, 2019; pp. 4932–4942.
26. Nagarajan, V.; Kolter, J.Z. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: New York City, NY, USA, 2019; pp. 11615–11626.
27. Tsutsumi, E.; Kinoshita, R.; Ueno, M. Deep-IRT with independent student and item networks. In Proceedings of the 14th International Conference on Educational Data Mining, EDM, Paris, France, 29 June–2 July 2021.
28. Corbett, A.T.; Anderson, J.R. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Model. User-Adapt. Interact.* **1994**, *4*, 253–278. [CrossRef]
29. González-Brenes, J.; Huang, Y.; Brusilovsky, P. Fast: Feature-aware student knowledge tracing. In Proceedings of the NIPS 2013 Workshop on Data Driven Education, Lake Taho, NV, USA, 9–10 December 2013.

30. Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L.J.; Sohl-Dickstein, J. Deep Knowledge Tracing. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: New York City, NY, USA, 2015; pp. 505–513.
31. Khajah, M.; Lindsey, R.V.; Mozer, M.C. How Deep is Knowledge Tracing? *arXiv* **2016**, arXiv:1604.02416.
32. Zhang, J.; Shi, X.; King, I.; Yeung, D.Y. Dynamic Key-Value Memory Network for Knowledge Tracing. In Proceedings of the 26th International Conference on World Wide Web, WWW '17, Perth, Australia, 3–7 May 2017; International World Wide Web Conferences Steering Committee: Geneva, Switzerland, 2017; pp. 765–774.
33. Vie, J.; Kashima, H. Knowledge Tracing Machines: Factorization Machines for Knowledge Tracing. *arXiv* **2018**, arXiv:1811.03388.
34. Pandey, S.; Karypis, G. A Self-Attentive model for Knowledge Tracing. In Proceedings of International Conference on Education Data Mining, Montreal, QC, Canada, 2–5 July 2019.
35. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
36. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; MESA Press: San Diego, CA, USA, 1993.
37. Hanson, B.A.; BÃ©guin, A.A. Obtaining a Common Scale for Item Response Theory Item Parameters Using Separate Versus Concurrent Estimation in the Common-Item Equating Design. *Appl. Psychol. Meas.* **2002**, *26*, 3–24. [CrossRef]
38. Hu, H.; Rogers, W.T.; Vukmirovic, Z. Investigation of IRT-Based Equating Methods in the Presence of Outlier Common Items. *Appl. Psychol. Meas.* **2008**, *32*, 311–333. [CrossRef]
39. Kolen, M.J.; Brennan, R.L. *Test Equating, Scaling, and Linking: Methods and Practices*, 2nd ed.; Springer: New York City, NY, USA, 2004.
40. Gonzãlez, J.; Wiberg, M. *Applying Test Equating Methods: Using R*; Springer: Berlin/Heidelberg, Germany, 2017.
41. Marco, G.L. Item Characteristic Curve Solutions to Three Intractable Testing Problems. *J. Educ. Meas.* **1977**, *14*, 139–160. [CrossRef]
42. Loyd, B.H.; Hoover, H.D. Vertical Equating Using the Rasch Model. *J. Educ. Meas.* **1980**, *17*, 179–193. [CrossRef]
43. Haebara, T. Equating Logistic Ability Scales by a Weighted Least Squares Method. *Jpn. Psychol. Res.* **1980**, *22*, 144–149. [CrossRef]
44. Stocking, M.L.; Lord, F.M. Developing a Common Metric in Item Response Theory. *Appl. Psychol. Meas.* **1983**, *7*, 201–210. [CrossRef]
45. Arai, S.; Mayekawa, S. A Comparison of Equating Methods and Linking Designs for Developing an Item Pool Under Item Response Theory. *Behaviormetrika* **2011**, *38*, 1–16. [CrossRef]
46. Sansivieri, V.; Wiberg, M.; Matteucci, M. A Review of Test Equating Methods with a Special Focus on IRT-Based Approaches. *Statistica* **2018**, *77*, 329–352. [CrossRef]
47. He, Y.; Cui, Z. Evaluating Robust Scale Transformation Methods With Multiple Outlying Common Items Under IRT True Score Equating. *Appl. Psychol. Meas.* **2020**, *44*, 296–310. [CrossRef]
48. Robitzsch, A. Robust Haebara Linking for Many Groups: Performance in the Case of Uniform DIF. *Psych* **2020**, *2*, 155–173. [CrossRef]
49. Robitzsch, A.; LÃŒdtke, O. Supplemental Material: A Review of Different Scaling Approaches under Full Invariance. *Psych. Test Assess. Model* **2020**, *62*, 233–279. [CrossRef]
50. Bock, R.D.; Zimowski, M.F. Multiple Group IRT. In *Handbook of Modern Item Response Theory*; Springer: Berlin/Heidelberg, Germany, 1997; pp. 433–448.
51. Jodoin, M.; Keller, L.; Swaminathan, H. A Comparison of Linear, Fixed Common Item, and Concurrent Parameter Estimation Equating Procedures in Capturing Academic Growth. *J. Exp. Educ.* **2003**, *71*, 229–250. [CrossRef]
52. Li, Y.; Tam, H.; Tompkins, L.J. A Comparison of Using the Fixed Common-Precalibrated Parameter Method and the Matched Characteristic Curve Method for Linking Multiple-Test Items. *Int. J. Test.* **2004**, *4*, 267–293. [CrossRef]
53. Shen, W.; Wang, X.; Bai, X.; Zhang, Z. DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3982–3991. [CrossRef]
54. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
55. Kilmen, S.; Demirtasli, N. Comparison of Test Equating Methods Based on Item Response Theory According to the Sample Size and Ability Distribution. *Procedia Soc. Behav. Sci.* **2012**, *46*, 130–134. [CrossRef]
56. Uysal, I.; Kilmen, S. Comparison of Item Response Theory Test Equating Methods for Mixed Format Tests. *Int. Online J. Educ. Sci.* **2016**, *8*, 1–11. [CrossRef]
57. Ueno, M. Animated agent to maintain learner's attention in e-learning. In *Proceedings of the E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2004*; Nall, J., Robson, R., Eds.; Association for the Advancement of Computing in Education (AACE): Washington, DC, USA, 2004; pp. 194–201.
58. Ueno, M. Data Mining and Text Mining Technologies for Collaborative Learning in an ILMS "Samurai". In Proceedings of the ICALT '04 Proceedings of the IEEE International Conference on Advanced Learning Technologies, Joensuu, Finland, 30 August–1 September 2004; pp. 1052–1053. [CrossRef]
59. Ueno, M. Intelligent LMS with an agent that learns from log data. In *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2005*; Richards, G., Ed.; Association for the Advancement of Computing in Education (AACE): Vancouver, BC, Canada, 2005; pp. 3169–3176.
60. George, A.C.; Robitzsch, A.; Kiefer, T.; Groß, J.; Ünlü, A. The R Package CDM for Cognitive Diagnosis Models. *J. Stat. Softw. Artic.* **2016**, *74*, 1–24. [CrossRef]

61. Ueno, M.; Miyazawa, Y. Probability Based Scaffolding System with Fading. In Proceedings of the Artificial Intelligence in Education—17th International Conference, AIED, Madrid, Spain, 21–25 June 2015; pp. 237–246._49. [CrossRef]
62. Ueno, M.; Miyazawa, Y. IRT-Based Adaptive Hints to Scaffold Learning in Programming. *IEEE Trans. Learn. Technol.* **2018**, *11*, 415–428. [CrossRef]
63. Wilson, K.H.; Karklin, Y.; Han, B.; Ekanadham, C. Back to the basics: Bayesian extensions of IRT outperform neural networks for proficiency estimation. In Proceedings of the 9th International Conference on Educational Data Mining, Raleigh, NC, USA, 29 June– 2 July 2016; Volume 1, pp. 539–544.
64. Ueno, M.; Pokpong, S. Computerized Adaptive Testing Based on Decision Tree. In Proceedings of the Advanced Learning Technologies (ICALT), 2010 IEEE Tenth International Conference, Sousse, Tunisia, 5–7 July 2010; pp. 191–193.
65. Ueno, M. Adaptive testing based on Bayesian decision theory. In Proceedings of the International Conference on Artificial Intelligence in Education, Memphis, TN, USA, 9–13 July 2013; pp. 712–716.