IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 June 2023, accepted 19 July 2023, date of publication 1 August 2023, date of current version 9 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3300375

RESEARCH ARTICLE

An Insight to Estimated Item Response Matrix in Item Response Theory

HIDEO HIROSE, (Member, IEEE)
Bioinformatics Center, Kurume University, Fukuoka 830-0011, Japan
e-mail: hirose_hideo@kurume-u.ac.jp

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 17H01842 and Grant 21K04558.

ABSTRACT This paper investigates the performance of item response theory based on distance criteria rather than likelihood criteria. For this purpose, the estimated item response matrix is introduced. This matrix is a reconstruction of the item response matrix using maximum likelihood estimates of the parameters in item response theory. Then the distance between the observed and estimated matrices can be determined using the Frobenius matrix norm. An approximated low-rank matrix can be generated from the observed item response matrix by singular value decomposition, and the distance between the observed and low-rank matrices can be obtained in the same way. By comparing these two distances, we can evaluate the performance of the estimated item response matrix comparable to the performance of an approximated low-rank matrix. Applying this comparison to actual examination data, it is found that the rank of the approximated low-rank matrix that is equivalent to the estimated item response matrix is very low when using matrices as training data. However, using test data, the predictive ability of item response theory seems high enough since the minimum distance between the approximated low-rank matrix and the observed item response matrix is approximately equal to or slightly less than the distance between the estimated item response matrix and the observed item response matrix. This fact has been first discovered by utilizing the estimated item response matrix defined here.

INDEX TERMS Computer based testing, estimated item response matrix, Frobenius matrix norm, item response theory, low-rank matrix, matrix completion, maximum likelihood estimation, observed item response matrix, singular value decomposition.

I. INTRODUCTION

Item response theory (IRT) (see [1], [2], [3], [4]) is a theory based on a statistical parametric model that simultaneously assesses abilities of examinees and difficulties of problems. Because of its versatility and reliability, IRT has been regarded as one of the standard methods for assessing examinee performance. For this reason, IRT is used in various official examinations, including the TOFLE. Configuring a matrix of examinee user rows and problem item columns with 0/1 valued responses (1 is success, 0 is failure), the maximum likelihood estimation method can obtain the estimates for IRT parameters and their confidence intervals. This matrix is called the *"observed item response matrix"* in this paper. In a word, IRT takes this matrix as input and outputs estimates.

The maximum likelihood estimators are known to be consistent and asymptotically efficient under certain conditions (see [5]); that is, no consistent estimators have lower asymptotic mean squared errors other than the maximum likelihood estimators. This means that the estimators perform best under the assumed mathematical model. Even though IRT is an ideal mathematical model defined on a certain support, IRT cannot strictly realize the real world. There could be a discrepancy between the model and the real. However, the likelihood criterion itself cannot be used to assess such a discrepancy. Other criteria may be used for such evaluation. In this paper, the distance criterion is used.

For this purpose, we propose to use the *"estimated item response matrix"* which is defined by the reconstructed item

The associate editor coordinating the review of this manuscript and approving it for publication was Sajid Ali.

response matrix using the maximum likelihood estimates for IRT parameters. As will be shown later, the performance in the maximum likelihood estimates using the observed item response matrix and the performance in the estimated item response matrix are considered to be approximately equivalent to each other. Then, we can measure the discrepancy between the real observed data and the IRT estimation result through the distance criterion. In this way the performance of IRT can be evaluated out of the specified mathematical model and its defined space.

Singular value decomposition (SVD) can generate approximated low-rank matrices from an observed item response matrix. To each approximated low-rank matrix, we can compute the distance between the approximated low-rank matrix and the observed item response matrix using the root mean squared error (RMSE). By comparing the RMSE between the estimated item response matrix and the observed item response matrix with the RMSE between the approximated low-rank item response matrix generated by SVD and the observed item response matrix, we can identify an approximately equivalent low-rank item response matrix to the estimated item response matrix. Using this rank and the RMSE, we can assess how far the estimated item response matrix locates from the real data; that is, we can know the position of the maximum likelihood estimates of IRT.

Amazingly, the rank of the equivalent approximate low-rank item response matrix relative to the estimated item response matrix turned out to be very low. Furthermore, the predictive accuracy of the estimated item response matrices is sufficient even for more complex matrices generated using the approximated higher-ranked item response matrices.

This fact is seen in not only one case, but in many CBT cases in undergraduate mathematics tests. Here, CBT refers to a computer based testing in which the response is 0 or 1. The purpose of this paper is to demonstrate this surprising fact. In this paper, as in recommender systems, the terms ''item'' and ''user'' are also referred to in the same way as the terms ''problem'' (or ''question'') and ''examinee''.

II. ESTIMATED ITEM RESPONSE MATRIX

A. MATHEMATICAL MODEL FOR ITEM RESPONSE THEORY

The standard IRT estimates proficiency parameters θ_i $(i = 1, \ldots, n)$ and problem parameters a_j, b_j, c_j $(j = 1, \ldots, m)$ simultaneously by using the observed item response matrix. Usually, this $n \times m$ size matrix consists of 0/1 valued elements δ_{ij}, with value 1 for (i, j) element corresponding to the case where examinee i solved question j correctly and with value 0 to the case where he/she did not solve it correctly. The observed item response matrix is expressed as $\Delta = (\delta_{ij})$.

Assuming that a logistic probability function p_{ij} of examinee i correctly answering question j is expressed such that

$$p_{ij}(\theta_i; a_j, b_j, c_j) = c_j + \frac{1 - c_j}{1 + \exp\{-1.7a_j(\theta_i - b_j)\}}, \\ = 1 - q_{ij}(\theta_i; a_j, b_j, c_j), \tag{1}$$

where θ_i is called the ability for examinee i and a_j, b_j, c_j are called the discrimination parameter, difficulty parameter, and pseudo-guessing parameter, respectively; q_{ij} is the probability that examinee i answers question j incorrectly.

B. MAXIMUM LIKELIHOOD PARAMETER ESTIMATION

Using the maximum likelihood estimation (MLE) method, the maximum likelihood estimates $\hat{\theta}_i, \hat{a}_j, \hat{b}_j, \hat{c}_j$ for parameters θ_i, a_j, b_j, c_j can be obtained by maximizing the likelihood function,

$$L_\Delta = \prod_{i=1}^{n} \prod_{j=1}^{m} \left(p_{ij}^{\delta_{ij}} \times q_{ij}^{1-\delta_{ij}} \right). \tag{2}$$

When only difficulty parameter b_j, in addition to parameter θ_i, is considered, such the model is called the Rasch model. Usually, the two-parameter model $(c_j = 0)$ is the standard when there are many choices in multiple-choice tests, and we will deal with this case in the following.

If we denote parameters θ_i, a_j, b_j together by Θ, then the estimation procedure is simply expressed as follows.

$$\Delta \rightsquigarrow \hat{\Theta}. \tag{3}$$

C. ESTIMATED ITEM RESPONSE MATRIX AND ESTIMATES

Substituting $\hat{\Theta}$ into p_{ij} in (1), we can obtain $\hat{p}_{ij}$ which is a continuous value in $[0, 1]$. The value $\hat{p}_{ij}$ is corresponding to the probability of answering the question correctly. It should be noted that $\hat{p}_{ij}$ can be regarded as $\hat{\delta}_{ij}$. Then, using $\hat{\Theta}$, a matrix can be constructed such that

$$\hat{\Theta} \rightsquigarrow \hat{\Delta}. \tag{4}$$

We call $\hat{\Delta} = (\hat{\delta}_{ij})$ the estimated item response matrix.

Consider the following likelihood function defined by

$$L_{\hat{\Delta}} = \prod_{i=1}^{n} \prod_{j=1}^{m} \left(p_{ij}^{\hat{\delta}_{ij}} \times q_{ij}^{1-\hat{\delta}_{ij}} \right). \tag{5}$$

Using the MLE method again, the maximum likelihood estimates $\check{\theta}_i, \check{a}_j, \check{b}_j$ for parameters θ_i, a_j, b_j can be obtained by maximizing the likelihood function $L_{\hat{\Delta}}$. Since each $\hat{\delta}_{ij}$ is composed from $\hat{\Theta}$, the parameter space for the likelihood function $L_{\hat{\Delta}}$ and that for L_Δ are considered to be almost the same. Then, the estimates by using L_Δ and those by using $L_{\hat{\Delta}}$ may become very close to each other. Using such an interesting phenomenon, the estimated item response matrix $\hat{\Delta}$ can be identified to an approximated low-rank matrix generated from SVD, as shown later.

III. SINGULAR VALUE DECOMPOSITION

A. SINGULAR VALUE DECOMPOSITION PROCEDURE

Assuming that $A = (a_{ij})$ is an $n \times m$ matrix. Then, $A^\mathsf{T}A$ becomes a symmetric $m \times m$ matrix, and AA^T becomes a symmetric $n \times n$ matrix, where A^T denotes the transpose of A. The eigen values and eigen vectors to these two matrices $A^\mathsf{T}A$ and AA^T are the same if they exist. We denote the eigen

values and eigen vectors to matrix $A^\mathsf{T}A$ as $\{\xi_1, \xi_2, \cdots, \xi_m\}$ and $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_m\}$. That is,

$$A^\mathsf{T}A\boldsymbol{v}_i = \xi_i \boldsymbol{v}_i. \tag{6}$$

Eigen values can be reordered such that $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_r > 0, \xi_{r+1} = \cdots = \xi_m = 0$, where r is the rank of $A^\mathsf{T}A$. Since $A^\mathsf{T}A$ is symmetric, eigen vectors can be made as orthonormal system. That is, $\boldsymbol{v}_i \cdot \boldsymbol{v}_j = I_{ij}$, where I_{ij} is the indicator function; i.e., $I_{ii} = 1$, and $I_{ij} = 0$ $(i \neq j)$. We make vector $\boldsymbol{u}_i$ by $\boldsymbol{u}_i = A\boldsymbol{v}_i/\sigma_i$, $(i \leq r)$, where $\sigma_i = \sqrt{\xi_i}$. In addition, if we produce matrices $U = (\boldsymbol{u}_i)$ and $V = (\boldsymbol{v}_i)$, then A can be expressed as $A = U\Sigma V^\mathsf{T}$, or equivalently, $A = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\mathsf{T}$. Here, Σ is a diagonal matrix using σ_i. This is the typical singular value decomposition (SVD) (see [6], [7], [8]).

B. GENERATING THE LOW-RANK MATRIX

We define A_k such that

$$A_k = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\mathsf{T}, \tag{7}$$

using the first k columns in the matrices of U and V. This procedure generates the *"approximated low-rank matrix"* A_k for A as shown below.

Theorem 1 (Eckart-Young [9]): 1) $rank(A_k) = k$
2) For any $n \times m$ matrix B, $(rank(B) \leq k)$,

$$||A - A_k||_F = \min_{B, rank(B) \leq k} ||A - B||_F = \Big(\sum_{i=k+1}^{m} \sigma_i^2\Big)^{1/2},$$

where $||\cdot||_F$ expresses the Frobenius matrix norm, i.e., $||(a_{ij})||_F = (\sum_{i,j} |a_{ij}|^2)^{1/2}$.

The theorem claims that A_k is best approximated to A among all the matrices with rank of less than $k + 1$ in the sense of matrix norm.

C. CONSTRUCTION OF THE LOW-RANK ITEM RESPONSE MATRIX

Applying the above method to the observed item response matrix Δ, the approximated low-rank item response matrix Δ_k can be constructed from Δ.

IV. DISTANCE CRITERION BETWEEN TWO MATRICES

The distance criterion of two equal-sized matrices $A = (a_{ij})$ and $B = (b_{ij})$ can be expressed by the RMSE(A, B) such that

$$\begin{aligned} \mathrm{RMSE}(A, B) &= \sqrt{\frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} (a_{ij} - b_{ij})^2} \\ &= \sqrt{\frac{1}{nm} (||A - B||_F)^2}. \end{aligned} \tag{8}$$

This is the case when all elements of A, B are occupied. In other words, matrices are complete. In such a situation, treating A as a prediction matrix for B may induce an overfitting phenomenon. This means that the prediction is made only for training data using the full matrix.

In order to measure accurate distance, test data must also be used. The matrix is then divided into two parts, one for training data and one for test data. In this situation, we have to deal with incomplete matrices.

A. RMSE FOR TEST DATA USING INCOMPLETE MATRIX TREATMENT

First, create two matrices S and T for training and test data, respectively. Here, S and T behave as if they were incomplete matrices. In the case of IRT and SVD, the RMSE of an incomplete matrix cannot be computed straightforwardly. However, in both cases, the algorithm for finding the RMSE of an incomplete matrix can be realized by an iterative method via the algorithm for the case of a complete matrix.

To define S and T, let $\Omega = (\omega_{ij})$ be a matrix and assume $\omega_{ij} = 0$ when δ_{ij} is used for training data and $\omega_{ij} = 1$ when δ_{ij} is used for test data. Then, $S = (s_{ij})$ and $T = (t_{ij})$ are defined such that

$$\begin{aligned} \Delta &= S + T, \\ s_{ij} &= \begin{cases} \delta_{ij} & (\omega_{ij} = 0) \\ 0 & (\omega_{ij} = 1), \end{cases} \\ t_{ij} &= \begin{cases} 0 & (\omega_{ij} = 0) \\ \delta_{ij} & (\omega_{ij} = 1) \end{cases} . \end{aligned} \tag{9}$$

S and T are actually complete matrices, but by incorporating this matrix Ω, they behave as if they were incomplete matrices.

B. IN THE CASE OF IRT

Usually, the element δ_{ij} takes a value such that when the question is answered, $\delta_{ij} = 1$ for success and $\delta_{ij} = 0$ for failure. However, we extended the value of δ_{ij} from a discrete value of 0/1 to a continuous value in $[0, 1]$ corresponding to the response level. We have also added a kind of matrix completion for cases where the value of element (i, j) is blank. Such a case corresponds to the case where examinee i was not working on question j, or where the value of response δ_{ij} was unknown. In the following, the algorithm for this procedure will be presented for training and test data sets. This algorithm is similar to [10], [11], and [12].

Algorithm (IRT):

1) set $z = 0$
2) set

$$\begin{aligned} &\begin{cases} s_{ij}^{(z)} = s_{ij} & (\omega_{ij} = 0) \\ s_{ij}^{(z)} = 0 & (\omega_{ij} = 1), \end{cases} \\ &\begin{cases} t_{ij}^{(z)} = 0 & (\omega_{ij} = 0) \\ t_{ij}^{(z)} \in [0, 1] & (\omega_{ij} = 1), \end{cases} \\ &W^{(z)} = S^{(z)} + T^{(z)} \end{aligned}$$

3) obtain $\hat{W}^{(z)} = \hat{S}^{(z)} + \hat{T}^{(z)}$ from $W^{(z)}$ using IRT

4)

$$\begin{cases} s_{ij}^{(z+1)} = s_{ij} & (\omega_{ij} = 0) \\ s_{ij}^{(z+1)} = 0 & (\omega_{ij} = 1), \end{cases}$$
$$\begin{cases} t_{ij}^{(z+1)} = 0 & (\omega_{ij} = 0) \\ t_{ij}^{(z+1)} = \hat{t}_{ij}^{(z)} & (\omega_{ij} = 1), \end{cases}$$
$$W^{(z+1)} = S^{(z+1)} + T^{(z+1)}$$

5) repeat 3) and 4) until $\hat{S}^{(z)}$ and $\hat{T}^{(z)}$ become stable
6) denote optimal $\hat{s}_{ij}$ as $\tilde{s}_{ij}$ and $\hat{t}_{ij}$ as $\tilde{t}_{ij}$

The RMSE for the training data case and the test data case are obtained such that

$$\mathrm{RMSE}(\tilde{S}, S) = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(1-\omega_{ij})(\tilde{s}_{ij}-s_{ij})^2}{\sum_{i=1}^{n}\sum_{j=1}^{m}(1-\omega_{ij})}},$$
$$\mathrm{RMSE}(\tilde{T}, T) = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_{ij}(\tilde{t}_{ij}-t_{ij})^2}{\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_{ij}}}. \quad (10)$$

C. IN THE CASE OF SVD

The RMSE for the test data of incomplete matrices in SVD can be obtained by modifying the algorithm of IRT. Let $\Delta_k = S_k + T_k$ be each low-rank matrix induced from Δ.

Algorithm (SVD):

1) for $k = 1, \ldots, k_{max}$, do 2) - 8) to each k; usually, $k_{max} = rank(\Delta)$
2) set $z = 0$
3) set

$$\begin{cases} s_{ij}^{(z)} = s_{ij} & (\omega_{ij} = 0) \\ s_{ij}^{(z)} = 0 & (\omega_{ij} = 1), \end{cases}$$
$$\begin{cases} t_{ij}^{(z)} = 0 & (\omega_{ij} = 0) \\ t_{ij}^{(z)} \in [0,1] & (\omega_{ij} = 1), \end{cases}$$
$$W^{(z)} = S^{(z)} + T^{(z)}$$

4) perform SVD to $W^{(z)}$ and obtain $U^{(z)}$, $V^{(z)}$, and $\Sigma^{(z)}$
5) set $W_k^{(z)} = \sum_{i=1}^{k} \sigma_i^{(z)} \boldsymbol{u}_i^{(z)} \boldsymbol{v}_i^{(z)\mathsf{T}}$ and rewrite $w_k^{(z)}$ as $\hat{w}^{(z)} = \hat{s}_{ij}^{(z)} + \hat{t}_{ij}^{(z)}$
6)

$$\begin{cases} s_{ij}^{(z+1)} = s_{ij} & (\omega_{ij} = 0) \\ s_{ij}^{(z+1)} = 0 & (\omega_{ij} = 1), \end{cases}$$
$$\begin{cases} t_{ij}^{(z+1)} = 0 & (\omega_{ij} = 0) \\ t_{ij}^{(z+1)} = \hat{t}_{ij}^{(z)} & (\omega_{ij} = 1), \end{cases}$$
$$W^{(z+1)} = S^{(z+1)} + T^{(z+1)}$$

7) repeat 4) - 6) until $\hat{S}^{(z)}$ and $\hat{T}^{(z)}$ become stable
8) denote stable $\hat{s}_{ij}$ as $\tilde{s}_{ij,k}$ and stable $\hat{t}_{ij}^{(z)}$ as $\tilde{t}_{ij,k}$

The RMSE for the training and test data cases for each k are obtained such that

$$\mathrm{RMSE}(\tilde{S}_k, S) = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(1-\omega_{ij})(\tilde{s}_{ij,k}-s_{ij})^2}{\sum_{i=1}^{n}\sum_{j=1}^{m}(1-\omega_{ij})}},$$

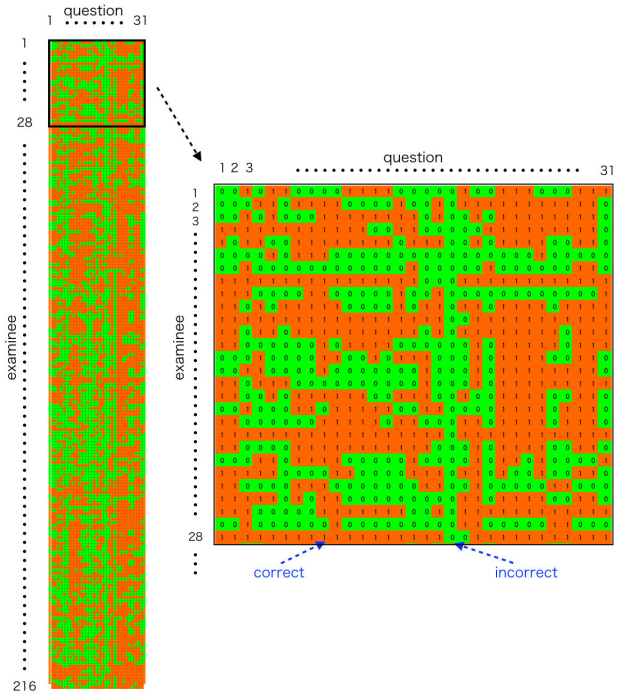

FIGURE 1. Observed item response matrix (case A).

$$\mathrm{RMSE}(\tilde{T}_k, T) = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_{ij}(\tilde{t}_{ij,k}-t_{ij})^2}{\sum_{i=1}^{n}\sum_{j=1}^{m}\omega_{ij}}}, \quad (11)$$

where $\Delta = S + T$, $\tilde{\Delta}_k = \tilde{S}_k + \tilde{T}_k$.

V. A TYPICAL EXAMINATION DATA CASE

A. OBSERVED ITEM RESPONSE MATRIX

As a typical data case study, we will use an examination data case derived from a mathematics midterm examination at a university. The number of examinees n is 216 and the number of questions m is 31. There are no missing data in this matrix. We name this example case A.

Fig. 1 shows the observed item response matrix of case A. On the right of the figure, responses of examinees from id 1 to id 28 are shown enlarged for clarity. This matrix consists of binary elements with 1 for correct answers and 0 for incorrect answers; this is denoted as $\Delta = (\delta_{ij})$.

B. ESTIMATED ITEM RESPONSE MATRIX

Applying the MLE method to the observed item response matrix Δ yields the maximum likelihood estimate $\hat{\Theta}$ for parameter Θ. Using this estimated value $\hat{\Theta}$, the estimated item response matrix $\hat{\Delta}$ can be constructed, as explained earlier. It should be noted that each $\hat{\delta}_{ij}$ becomes the maximum likelihood estimate for δ_{ij}.

The figure on the left in Fig. 2 shows $\hat{\Delta}$ for case A. Comparing $\hat{\Delta}$ with Δ in Fig. 1, we can roughly imagine the original observed item response matrix Δ from $\hat{\Delta}$. However, at first glance, this approximation appears to be inaccurate.

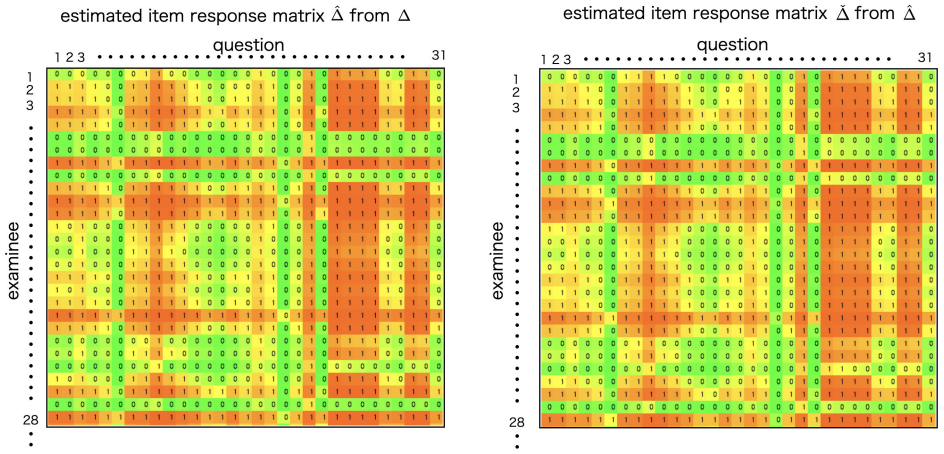

FIGURE 2. Two kinds of estimated item response matrices (case A).

To see if the distance between the observed item response matrix Δ and the estimated item response matrix $\hat{\Delta}$ is large or small, we have computed the RMSE($\hat{\Delta}$, Δ). It was 0.3915, and this indicates that the distance between an observed δ_{ij} and its estimated value $\hat{\delta}_{ij}$ lies on average around 0.3915. Intuitively, this value does not seem small. This is consistent with what we indicated above. One might think that IRT is not working well. However, it will be understood that IRT performs very well by comparing the RMSE of IRT and that of SVD.

Before we do that, we need to see if it makes sense to compare $\hat{\Theta}$ performance with $\hat{\Delta}$ performance. As mentioned earlier, we can again obtain the maximum likelihood estimates $\check{\Theta}$ using $\hat{\Delta}$, and the corresponding estimated item response matrix $\check{\Delta}$ can be obtained. This is shown in Fig. 2 on the right. These two matrices $\check{\Delta}$ and $\hat{\Delta}$ are very similar to each other. Actually, the RMSE between the two estimated item response matrices RMSE($\check{\Delta}$, $\hat{\Delta}$) is 0.0275, which is considered to be small. In addition, the value of $\cos(\check{\Theta}, \hat{\Theta}) = 0.992$ indicates a close similarity between these two estimates. Therefore, the performance of the maximum likelihood estimates $\hat{\Theta}$ can be regarded as approximately equivalent to the performance of the estimated item response matrix $\hat{\Delta}$.

Fig. 3 illustrates the diagram of procedure in comparing the maximum likelihood estimates $\hat{\Theta}$ and the estimated item response matrix $\hat{\Delta}$. We ultimately intend to compare the performances between $\hat{\Theta}$ and Δ_k derived from SVD. However, this cannot be done straightforwardly. Instead, we consider to use $\hat{\Delta}$ as a substitute for $\hat{\Theta}$. Since $\check{\Theta}$ and $\hat{\Theta}$ are approximately equivalent, we can assume that $\hat{\Theta}$ can be mapped back to $\hat{\Delta}$.

C. SINGULAR VALUE DECOMPOSITION

Singular value decomposition for the observed item response matrix Δ of case A resulted in the singular values and the RMSE(Δ_k, Δ) between the observed item response matrix Δ and the approximated low-rank matrices Δ_k. These singular values and RMSE are shown in Fig. 4 on the left and right, respectively. The RMSE between the observed item response matrix and the estimated item response matrix using IRT, RMSE($\hat{\Delta}$, Δ), is superimposed in a straight line parallel to the horizontal axis in Fig. 4 on the right.

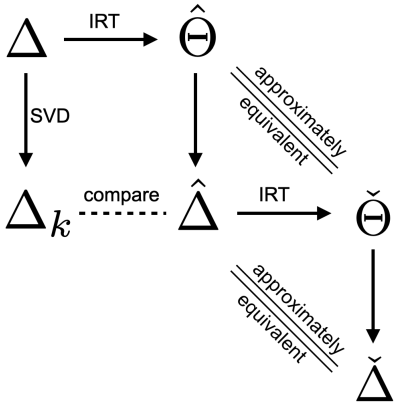

FIGURE 3. Diagram of procedure in comparing the estimates and the matrix.

Looking at the singular values, we see that only the largest singular value is outstanding and other singular values are rather small. In contrast to this, RMSE(Δ_k, Δ) is approximately linearly decreasing as k increases, and finally RMSE(Δ_{31}, Δ) = 0, which means that Δ_{31} is exactly the same as Δ.

This fact indicates that $\sigma_1, \boldsymbol{u}_1, \boldsymbol{v}_1^{\mathsf{T}}$ play an important role in determining the RMSE. Here, σ_1 is the largest singular value, $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ are the corresponding vectors.

Fig. 5 shows approximated low-rank matrices Δ_1, Δ_2, Δ_3, Δ_{10} for case A. Comparing $\hat{\Delta}$ in Fig. 2 with Δ_1 or Δ_2 in Fig. 5, all three appear to be similar, as Δ_1 or Δ_2 is similar to $\hat{\Delta}$. This is consistent with the value of RMSE($\hat{\Delta}$, Δ) being between the values of RMSE(Δ_1, Δ) and RMSE(Δ_2, Δ).

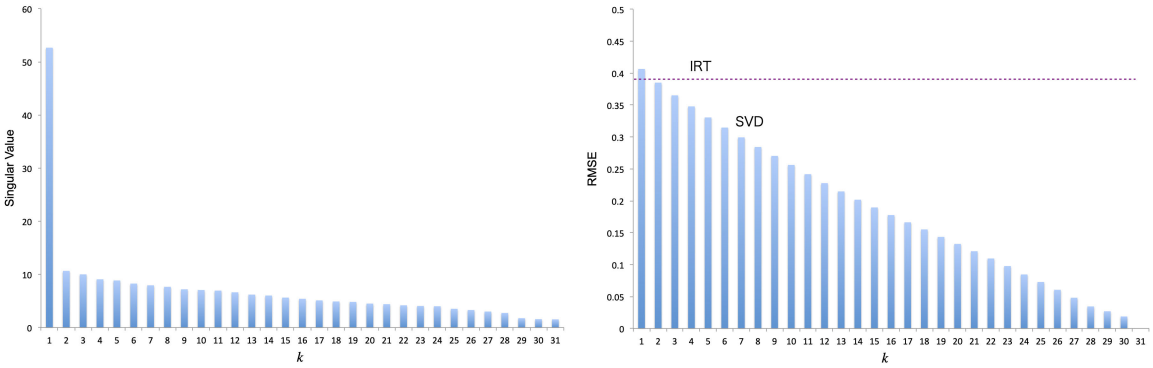

FIGURE 4. Singular value and RMSE for the observed item response matrix (case A).

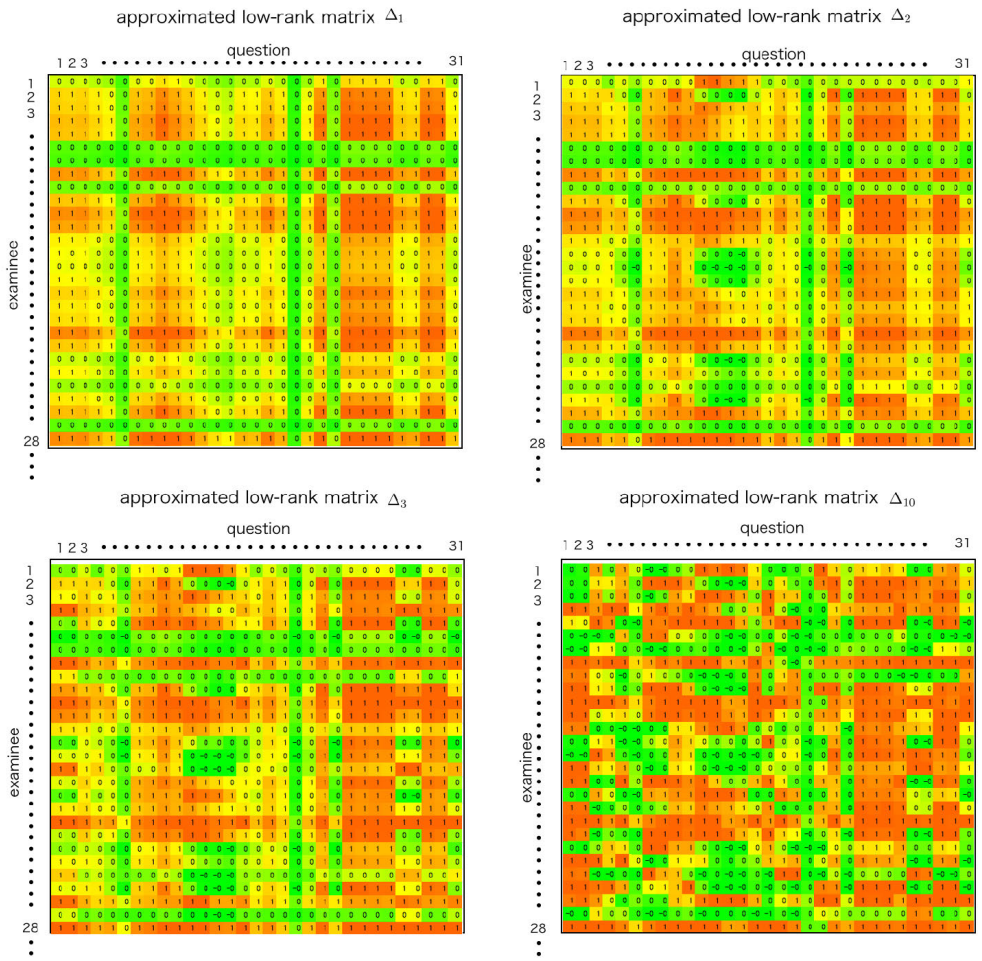

FIGURE 5. Approximated low-rank matrices for k=1, 2, 3, 10 (case A).

However, this is just an illusion because the matrix was treated completely as training data. To avoid overfitting, the RMSE should be evaluated using test data.

D. EVALUATION USING TEST DATA

Various methods have been proposed to overcome the effects of overfitting, including holdout and 10-fold cross validation. In this paper, we use a method similar to the latter, but with a different selection of training and test data. First, 10% of the elements from the original matrix are randomly selected and used as test data T, and the remaining 90% of the elements are used as training data S. Then, RMSE($\tilde{S}, S$) for training data and RMSE($\tilde{T}, T$) for test data are obtained in IRT, and RMSE($\tilde{S}_k, S$) for training data and RMSE($\tilde{T}_k, T$) for test

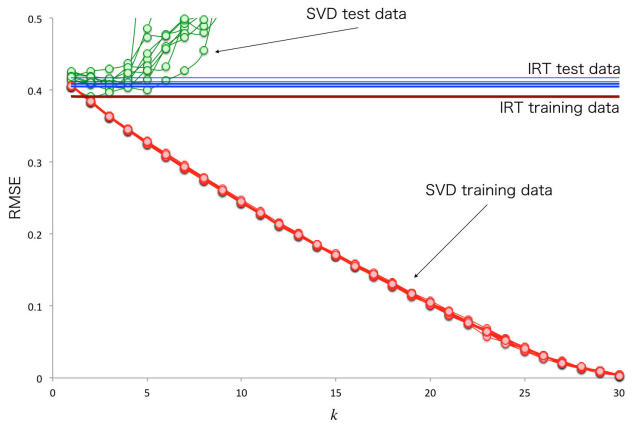

FIGURE 6. RMSE for the test data and the training data via SVD and IRT (case A).

data are obtained in SVD. This is repeated (bootstrapped) 10 times, and mean and standard deviation for RMSE are obtained.

Fig. 6 shows RMSE($\tilde{S}_k, S$) for training data and RMSE($\tilde{T}_k, T$) for test data as a function of k used in SVD. In the figure, 10 bootstrapped results are shown for SVD; at the same time, 10 IRT bootstrapped results, RMSE($\tilde{S}, S$) and RMSE($\tilde{T}, T$), are superimposed in straight lines parallel to the horizontal axis. The figure presents typical monotonically decreasing curves for training data and U-shaped curves for test data.

The figure tells us very interesting points as follows:

1) In general, RMSE($\tilde{T}_k, T$) in SVD shows a U-shaped curve as a function of k. In this case, however, the values of RMSE($\tilde{T}_k, T$) in SVD show rather flat curves when $k \leq 4$, and they increase when $k \geq 5$.
2) In particular, RMSE($\tilde{T}, T$) of IRT appears to be comparable to RMSE($\tilde{T}_k, T$) in SVD for $1 \leq k \leq 3$.

According to this, the estimated item response matrix $\hat{\Delta}$ can be regarded as an approximated low-rank matrices Δ_k with a small value of k in terms of RMSE distance criterion. In other words, the performance of the maximum likelihood estimates for IRT parameters is comparable to that in Δ_k when the value of k is small. However, this property is only obtained from one example, and it is necessary to check whether this property holds in other cases.

E. 42 EXAMINATION CASES

To make sure that the above mentioned property holds true for other examination cases, we collected 42 test cases, including [13]. Table 1 shows the subjects and matrix sizes of the 42 examination cases. All examinations were administered at universities and the examinees were undergraduate students. Subjects included probability, statistics, ordinary differential equations, calculus, and linear algebra. In all examinations, answers were given as discrete values of 0/1 (1 for correct answers and 0 for incorrect answers).

TABLE 1. Subjects and matrix size in 42 examination cases.

id	subject	n	m	*id*	subject	n	m
1	PS	44	14	22	LA	132	84
2	PS	41	19	23	LA	177	49
3	PS	34	17	24	LA	142	45
4	P	97	32	25	LA	46	39
5	S	57	15	26	LA	39	45
6	S	75	14	27	LA	181	45
7	C	40	19	28	LA	229	84
8	PS	44	15	29	C	1131	77
9	PS	72	21	30	LA	1101	84
10	ODE	41	13	31	C	215	36
11	ODE	49	25	32	LA	47	39
12	PS	54	21	33	C	209	36
13	C	70	26	34	C	868	6
14	C	9	16	35	C	215	36
15	C	45	42	36	LA	585	84
16	LA	36	39	37	LA	39	39
17	LA	566	6	38	C	209	31
18	C	66	33	39	C	209	67
19	S	97	12	40	C	216	31
20	C	76	30	41	LA	585	49
21	LA	132	49	42	C	145	34

PS: probability & statistics, P: probability, S: statistics, ODE: ordinary differential equations, C: calculus, LA: linear algebra

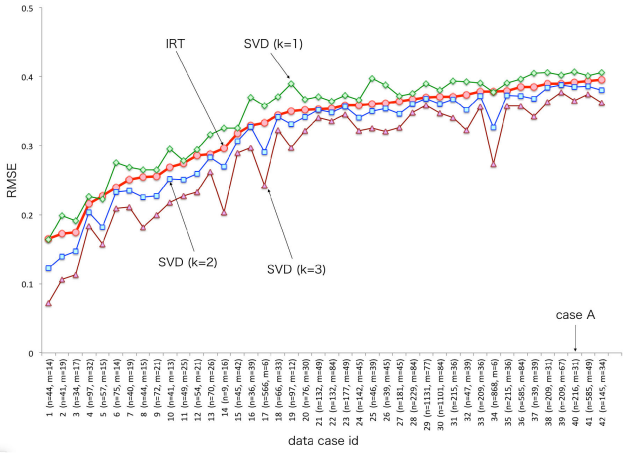

FIGURE 7. RMSE using IRT and SVD with $k = 1, 2, 3$ for 42 complete matrix using full element data.

As explained above, it would be sufficient to focus on RMSE($\tilde{T}_k, T$) for smaller k to compare RMSE($\tilde{T}_k, T$) with RMSE($\hat{\Delta}, \Delta$). For the sake of brevity, we first investigate RMSE($\hat{\Delta}_k, \Delta$) for lower rank of k and RMSE($\hat{\Delta}, \Delta$).

Fig. 7 shows the RMSE($\hat{\Delta}, \Delta$) for the 42 examination cases. In the figure, the cases *id* shown on the horizontal axis are arranged in ascending order of magnitude of the RMSE($\hat{\Delta}, \Delta$) shown on the vertical axis for clarity. Also shown are the RMSE(Δ_k, Δ) ($k = 1, 2, 3$) for each case *id*. The figure shows that RMSE(Δ_2, Δ) < RMSE($\hat{\Delta}, \Delta$) holds in all cases. Using the complete training data, we confirmed that the properties described in the previous section hold for all 42 test cases. For the sake of simplicity, we have chosen four cases to see if this property holds true even with the test data.

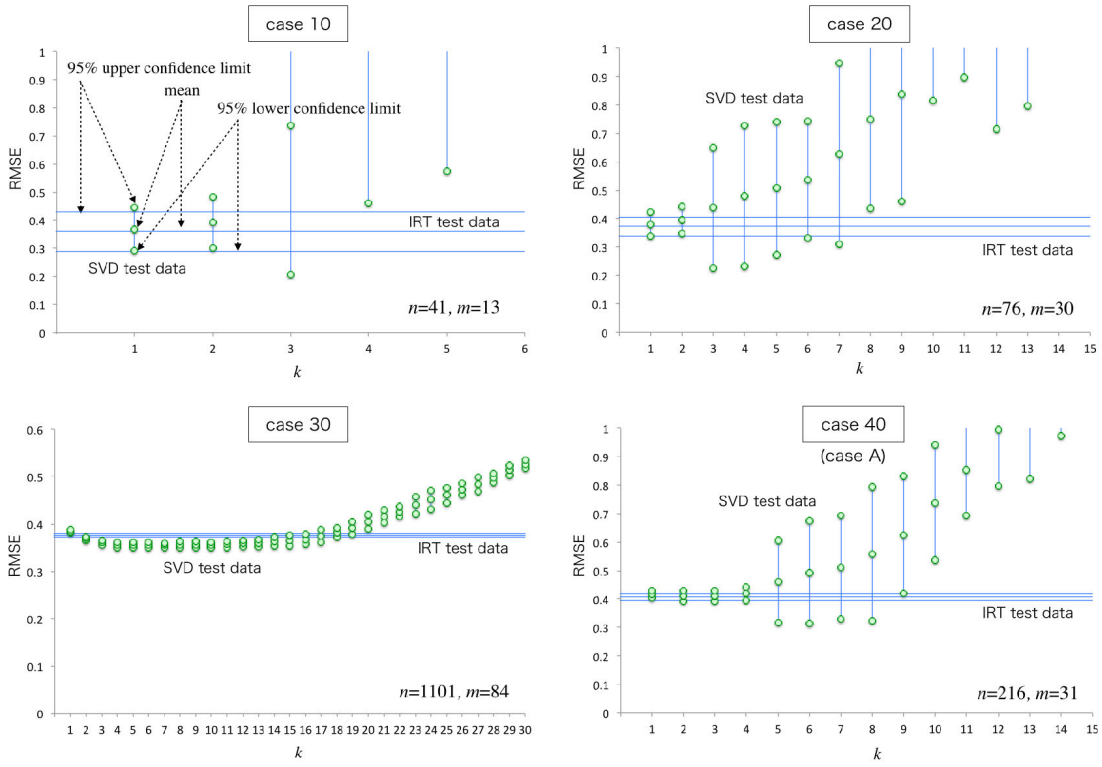

FIGURE 8. RMSE for the test data and the training data via MD and SVD (4 cases).

F. FOUR EXAMINATION CASES AMONG 42 CASES

From the 42 data cases, we selected four cases, including case A, to verify whether the RMSE of the test case in IRT is close to the RMSE of the low-rank matrix in SVD. These are cases 10, 20, 30, and 40 shown in Table 1. The values of RMSE($\tilde{\Delta}$, Δ) are ordered in ascending order.

Fig. 8 shows the RMSE for the test data using IRT and SVD. Circles placed at each k indicate the lower 95% confidence limit, mean, and upper 95% confidence limit of the RMSE from 10 bootstrapped computations in SVD. The IRT ones are also presented by straight lines parallel to the horizontal axis.

Looking at the figure, we see the following:

1) In all cases, upper and lower 95% confidence limits for the test data using IRT are almost the same as those using SVD with $k = 1, 2$. This property is the same as shown in using the training data when $k = 1$ or $k = 2$.
2) The smallest mean value of RMSE($\tilde{T}_k, T$) using SVD are obtained when the rank of the corresponding approximated low-rank matrix is small compared to the rank of matrix Δ, as shown in Table 2. This rank is called the optimum rank, and denoted by k_opt.
3) In all cases, 95% confidence limits for the test data using IRT are not so different from those using SVD for k_opt.
4) However, when the estimated item response matrix $\hat{\Delta}$ is quite complex, as in case 30, there is still room to develop a more accurate method of predicting examinee proficiency.

TABLE 2. Optimum rank in SVD and mean RMSE.

case *id*	rank of Δ	k_opt	RMSE1	RMSE2
10	12	1	0.3681	0.3591
20	30	1	0.3810	0.3726
30	77	7	0.3555	0.3771
40	31	3	0.4115	0.4068

RMSE1: mean of RMSE($\tilde{T}_{k\_opt}, T$),
RMSE2: mean of RMSE($\tilde{T}, T$)

In general, the predictive ability of IRT seems high enough since the RMSE($\tilde{T}_{k\_opt}, T$) using SVD is comparable to the RMSE($\hat{T}, T$) using IRT in moderate sized observed item response data. However, there is still room to develop a more accurate method of predicting examinee proficiency when the estimated item response matrix Δ becomes more complex. This fact has been first discovered by utilizing the estimated item response matrix defined here, and indicates the significance of the literature seen in [14], [15], [16], [17], [18], [19], [20], and [21].

VI. CONCLUDING REMARKS

IRT outputs the maximum likelihood estimates for parameters of the IRT model from the observed item response matrix. Using the estimates, the item response matrix can

be reconstructed. This is called the estimated item response matrix. Then the distance between the observed and estimated matrices can be determined using the Frobenius matrix norm. SVD generates an approximated low-rank matrix from the observed item response matrix, and the distance between the observed and low-rank matrices can be obtained in the same way.

By comparing these two distances, we can evaluate the performance of the estimated item response matrix comparable to the performance of an approximated low-rank matrix. In such a way, the performance of IRT can be evaluated.

Applying this method to actual examination data, it is found that the rank of the approximated low-rank matrix that is equivalent to the estimated item response matrix is very low when using matrices as training data. However, using test data, the predictive ability of IRT seems high enough since the minimum distance between the approximated low-rank matrix and the observed item response matrix is approximately equal to or slightly less than the distance between the estimated item response matrix and the observed item response matrix.

In general, the predictive ability of IRT seems high enough in moderate sized observed item response data. However, there is still room to develop a more accurate method of predicting examinee proficiency when the observed item response matrix becomes more complex. This fact has been first discovered by utilizing the estimated item response matrix defined here.

REFERENCES

[1] R. de Ayala, *The Theory and Practice of Item Response Theory*. New York, NY, USA: Guilford Press, 2009.

[2] F. B. Baker and S.-H. Kim, *Item Response Theory: Parameter Estimation Technique*, 2nd ed. New York, NY, USA: Marcel Dekker, 2004.

[3] R. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of Item Response Theory*. Newbury Park, CA, USA: Sage, 1991.

[4] W. J. van der Linden, *Handbook of Item Response Theory*. London, U.K.: Chapman & Hall, 2016.

[5] M. G. Kendall and A. Stuart, *Advanced Theory of Statistics*. New York, NY, USA: Macmillan Publishers, 1983.

[6] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD, USA: Johns Hopkins Univ. Press, 2012.

[7] G. Strang, "Multiplying and factoring matrices," *Amer. Math. Monthly*, vol. 125, no. 3, pp. 223–230, Mar. 2018.

[8] G. Strang, *Introduction to Linear Algebra*. Wellesley, MA, USA: Wellesley-Cambridge Press, 2021.

[9] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, Sep. 1936.

[10] T. Sakumura, T. Kuwahata, and H. Hirose, "An adaptive online ability evaluation system using the item response theory," in *Proc. Educ. e-Learn.*, 2011, pp. 51–54.

[11] H. Hirose and T. Sakumura, "Item response prediction for incomplete response matrix using the EM-type item response theory with application to adaptive online ability evaluation system," in *Proc. IEEE Int. Conf. Teaching, Assessment, Learn. Eng. (TALE)*, Aug. 2012, pp. T1A-6–T1A-10.

[12] T. Sakumura and H. Hirose, "Making up the complete matrix from the incomplete matrix using the EM-type IRT and its application," *Trans. Inf. Process. Soc. Jpn.*, vol. 7, no. 2, pp. 17–26, 2014.

[13] H. Hirose, "Meticulous learning follow-up systems for undergraduate students using the online item response theory," in *Proc. 5th IIAI Int. Congr. Adv. Appl. Informat. (IIAI-AAI)*, Jul. 2016, pp. 427–432.

[14] D. Reckase, *Multidimensional Item Response Theory*. New York, NY, USA: Springer, 2011.

[15] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 505–513.

[16] J.-J. Vie, "Deep factorization machines for knowledge tracing," in *Proc. 13th Workshop Innov. Use NLP Building Educ. Appl.*, 2018, pp. 370–373.

[17] J.-J. Vie and H. Kashima, "Knowledge tracing machines: Factorization machines for knowledge tracing," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, pp. 750–757.

[18] N. Thai-Nghe, L. Drumond, T. Horva, A. Krohn-Grimberghe, A. Nanopoulos, and L. Schmidt-Thieme, "Factorization techniques for predicting student performance," in *Educational Recommender Systems and Technologies: Practices and Challenges*. Hershey, PA, USA: IGI Global, 2011, pp. 129–153.

[19] M. Sweeney, J. Lester, H. Rangwala, and A. Johri, "Next-term student performance prediction: A recommender systems approach," *J. Educ. Data Mining*, vol. 8, no. 1, pp. 22–51, 2016.

[20] M. Khajah, Y. Huang, J. P. Gonzlez-Brenes, M. C. Mozer, and P. Brusilovsk, "Integrating knowledge tracing and item response theory: A tale of two frameworks," in *Proc. 4th Int. Workshop Personalization Approaches Learn. Environ.*, 2014, pp. 7–15.

[21] K. H. Wilson, X. Xiong, M. Khajah, R. V. Lindsey, S. Zhao, Y. Karklin, E. G. Van Inwegen, B. Han, C. Ekanadham, J. E. Beck, N. Heffernan, and M. C. Mozer, "Estimating student proficiency: Deep learning is not the panacea," in *Proc. Workshop Mach. Learn. Educ., Neural Inf. Process. Syst.*, 2016, pp. 1–8.

HIDEO HIROSE (Member, IEEE) received the bachelor's degree in mathematics from Kyushu University, in 1977, and the Dr. (Eng.) degree from Nagoya University, in 1988.

He was with Takaoka Electric Manufacturing Company Ltd., from 1977 to 1995, and the Vice Research Director, from 1988 to 1995. He was a Professor with Hiroshima City University, from 1995 to 1998, Kyushu Institute of Technology, from 1995 to 2015, and Hiroshima Institute of Technology, from 2015 to 2020. Since 2020, he has been a Visiting Professor with Kurume University and Chuo University. His research interests include reliability engineering, machine learning, data science, and educational technology.

Dr. Hirose is a member of American Statistical Association (ASA) and Japanese Academic Society. He was a recipient of the 2006 IEEJ Distinguished Paper Award, the 2009 CIEC Best Paper Award, and the 2010 International Conference on Computers and Advanced Technology in Education Best Paper Award.