Evaluation of the accuracy of ordinal classifications using item response theory

Mayu Ogawa¹, Mirai Tanaka², Xiao-Nan Lu¹, and Tomomichi Suzuki¹

¹Department of Industrial Administration, Tokyo University of Science, 2641 Yamazaki, Noda, Chiba, 278-8510, Japan, JSQC
²The Institute of Statistical Mathematics, 10-3 Midori, Tachikawa, Tokyo, Japan.

*contact author’s e-mail address: 7417605@ed.tus.ac.jp

Abstract:
It is very important to guarantee measurement results. For quantitative data, a methodology for obtaining accurate measurements in terms of both trueness and precision has been established according to ISO 5725. On the other hand, no methodology has been established for qualitative data. Moreover, in the previous study, the data are mostly supposed to be binary, rather than multinomial.

The purpose of this study is to validate a method of evaluating a measurement accuracy for ordinal-categorical (multinomial) data using the item response theory (IRT) developed in the fields of psychology and ability evaluation. The IRT model can estimate parameters indicating item difficulty and item discrimination and can evaluate whether the items can be classified appropriately according to the ability of the examinees. Two indicators of accuracy are proposed: correct-ordering probability and consistent-classification probability. We provide the characteristics of this method. The measurement data under different conditions were prepared, applied to the IRT model, and used to calculate the indicators. As a result, the two indicators could show appropriate values in most datasets, but some conditions did not work.

Keywords
Repeatability and reproducibility, Ordinal categorical data, Multinomial data, Measurement

1. Introduction
It is important to evaluate measurement methods when we perform quality assurance of measurement results. For example, during visual inspection of industrial products, measurement results from a plurality of inspectors and those repeatedly measured by the same inspector should be the same or similar. In reality, however, identical results are not necessarily produced. Therefore, it is important to evaluate variations in measurement results.

ISO 5725-1 (1994) and ISO 5725-2 (1994) are international standards for performance evaluation of measurement methods. Currently, ISO 5725 has an established methodology for determining the accuracy of quantitative data taking consecutive values. Accuracy represents how close the measurement results and true values match in terms of trueness and precision. On the other hand, for qualitative data taking discrete values, while many methods for evaluating the accuracy of binary data have been studied, there have been few studies on the accuracy of multiple-valued data such as nominal-scale and ordinal-scale data. Ordinal-scale data are employed in the various fields such as education and industry, for example:

• School record (excellent, good, passing, failing, for example)
• Item quality scales (low, medium, high, for example)
• Pathological item of animal experiment
Dispersion measure for ordinal data (each datum measured according to levels) was suggested by Blair and Lacy [3]. ORDANOVA, a model proposed by Bashkansky et al. (2012), describes how to analyze ordinal-scale data.

Item response theory (IRT) is said to be modern test theory, the basic idea was given by Lord (1980). It was developed in the fields of pedagogy and psychology as a test theory; it estimates parameters indicating item threshold and item discrimination and can evaluate whether items can be appropriately classified according to the ability of examinees.

In this paper, we evaluate whether IRT is a suitable method for evaluating the accuracy of measurements for ordinal-categorical (i.e., multinomial) data. There are various models of IRT. The multivalued IRT model has a graded response model proposed by Samejima (1968) and a partial credit model (PCM) proposed by Masters (1982). A generalized partial score model is a generalized partial credit model (GPCM) which is an extension of PCM was proposed by Muraki (1992). De Mast and Van Wieringen (2010) proposed a model for application to ISO 5725 using GPCM; it has two indicators of accuracy. Here we conduct a simulation using this model and its indicators. In the simulation, we prepare measured value data and examine whether the accuracy can be evaluated by the proposed IRT model and the indicators. Furthermore, we provide the characteristics of this model and indicators. The GPCM and the simulation are described in Sections 3 and 4, respectively.

2. Accuracy of measurement in ISO 5725

2.1 Accuracy of measurement of quantitative data

According to ISO 5725, measurement result accuracy can be gauged in terms of two indicators, namely, trueness and precision. Trueness is a bias that shows how well the average value and the true value of the measurement result agree, whereas precision is a variation that indicates the extent to which the measurement results match. Using this standard, we conduct collaborative-assessment experiments that measure the same sample in several laboratories using the same method to evaluate the measuring ability of each laboratory.

In addition, when we evaluate the measurement precision, we do so in terms of repeatability and reproducibility. Repeatability is the accuracy of measurement under repeated conditions (i.e., within the shortest time period using the same sample, method, laboratory, experimenter, and device), whereas reproducibility is the accuracy of measurement under reproduced conditions (i.e., independent measurement results obtained using the same samples and measurement method, but different laboratories, experimenters, and devices). Therefore, repeatability represents the variation inside a laboratory, whereas reproducibility represents the variation between laboratories.

In order to estimate the accuracy of the measurement method, the basic model of the measurement result $y$ is denoted by

$$y = m + B + e.$$ 

Here $m$ is the level or true value of the sample, $B$ is the laboratory component, and $e$ is the random error associated with each measurement. When $m$ is a true value, it is necessary to consider the bias, $B$, to be constant in measurements made under repeated conditions; however, $B$ takes different values for measurements made under reproduced conditions. The variance, $\sigma_e^2$, is called inter-room variance and denoted by

$$\sigma_e^2 = V(B).$$

The standard deviation of $e$ is taken as the repeatability standard deviation, $\sigma_r$, and the standard deviations of $e$ and $B$ are taken as the reproducibility standard deviation, $\sigma_R$. These are defined, respectively, as

$$\sigma_r = \sqrt{V(e)}$$

$$\sigma_R = \sqrt{V(B)}$$
and

\[ \sigma_R = \sqrt{\sigma_L^2 + \sigma_R^2}. \]

### 2.2 Measurement accuracy in qualitative data

Currently, neither concepts nor models have been established with regard to the accuracy of qualitative data. For ordinal-categorical data, it has only information on order, not equidistant. They can be handled in the same way as quantitative data, but are supposed to be equidistant. This deficiency is partly due to the inability to assume a normal distribution, as with quantitative data. In a previous study, an IRT model was proposed as a method for evaluating the measurement accuracy of order-categorical data by De Mast and Van Wieringen (2010).

### 3. IRT

#### 3.1 IRT model

While grading a test, there are cases wherein a score is not only given in terms of correct/incorrect answers but also in terms of 0, 1, 2, or 3 points. In such cases, the model applied to the multivalued data is called a multivalued IRT model. A representative model is the GPCM, which is represented as

\[
P_{jh}(\theta) = \frac{\exp\left(a_j \sum_{m=0}^{h} (\theta - b_{jm})\right)}{\sum_{h=0}^{H} \exp\left(a_j \sum_{m=0}^{h} (\theta - b_{jm})\right)},
\]

where \( j = 0, 1, 2, \ldots, J \) denotes a test item, \( h = 0, 1, 2, \ldots, H \) denotes a partial score, and \( \theta \) denotes an ability of the examinee. We assume that \( \theta \) follows a standard normal distribution. \( P_{jh}(\theta) \) represents the probability that an examinee of ability \( \theta \) obtains score \( h \) on item \( j \). The item-discrimination parameter \( a_j \) represents the ability of item \( j \) to classify examinees with different abilities by partial score, and the item-threshold parameter \( b_{jm} \) represents the boundary of the ability value of the examinee when their ability is classified by score.

A graph obtained by GPCM as a function of \( \theta \) is called the category-probability curve of item \( j \). The horizontal axis is the ability value \( \theta \) of the examinee and the vertical axis is the probability \( P_{jh}(\theta) \). Figure 1 is a categorical-probability curve in item \( j \) when the partial score is 0, 1, 2, or 3 points. The item-discrimination parameter \( a_j \) is the slope of the tangent at \( \theta = b_{jm} \) and the item-threshold parameter \( b_{jm} \) is the examinee's ability value \( \theta \) when the curves of the adjacent scores \( h + 1 \) and \( h \) intersect.

![Category-probability curve for item j](image)

**Figure 1. Category-probability curve for item j**

#### 3.2 Application of IRT to ISO 5725
In ISO 5725, it is assumed that the measurement values differ for a sample due to variation between and within laboratories. On the other hand, in IRT, it is assumed that the evaluation of the test differs for examinee groups due to variation between and within test items.

Comparing ISO 5725 and IRT, we believe that laboratories and repetition correspond to items and the number of examinees, respectively. Therefore, for the IRT applied to ISO 5725, it is assumed that the measurement values differ over the whole sample due to variation of the samples between and within laboratories. It is assumed that the partial score corresponds to the categories of classification.

The GPCM model proposed by De Mast and Van Wieringen (2010) for application to ISO 5725 is represented by

\[
q_j(h|x) = P(Y_{ij} = h|X_i = x) = \frac{\exp \left( \alpha_j \sum_{m=0}^{h} (x - \delta_{jm}) \right)}{\sum_{h=0}^{H} \exp \left( \alpha_j \sum_{m=0}^{h} (x - \delta_{jm}) \right)},
\]

where \( j = 0, 1, 2, \ldots J \) denotes a laboratory, \( h = 0, 1, 2, \ldots H \) denotes a category, and \( x \) denotes a true value of the sample. We assume that \( x \) follows the standard normal distribution. \( q_j(h|x) \) represents the probability that the sample of true value \( x \) is classified as category \( h \) in laboratory \( j \). \( Y_{ij} \) is the measured value of sample \( i \) in laboratory \( j \), and \( X_i \) is the true value of the sample \( i \). \( Y_{ij} \) is a random variable that follows a multinomial distribution and \( X_i \) is a random variable that follows the standard normal distribution, \( N(0, 1^2) \). The item-discrimination parameter \( \alpha_j \) represents the ability of laboratory \( j \) to classify samples with different true values by category, and the threshold parameter \( \delta_{jm} \) represents the boundary of the true value of the sample, that is classified as categories \( m, h \).

Figure 2 is the category-probability curve in laboratory \( j \) when the category is 0, 1, 2, 3. The item-discrimination parameter \( \alpha_j \) is the slope of the tangent line at \( \theta = \delta_{jm} \) and the item-threshold parameter \( \delta_{jm} \) is the true value of \( x \) when the curves of the adjacent categories \( h - 1 \) and \( h \) intersect.

3.3 Measurement accuracy in IRT

As a scale of measurement accuracy, we use the correct-ordering probability \( \rho \) and the consistent-classification probability \( \pi \). These were proposed by De Mast and Van Wieringen (2010). \( \rho \) represents whether ordering is correct between and within laboratories; \( \pi \) represents whether the threshold (category boundary) in GPCM is consistent between and within laboratories.

The indicator \( \rho_j^w \) (\( w \) means within) expresses the extent of matching between the ordering of the measurement value and the true value within laboratory \( j \). \( \rho_j^w \) is defined as eq (3).

\[
\rho_j^w = P(Y_{ij} \leq Y_{uj}|X_i \leq X_u)
\]

\( \rho_j^w \) is a value obtained by rescaling \( \rho_j^w \) to [0, 1]. It is calculated \( \rho_j^w = (\rho_j^w - (h + 1)/2h)(1 - (h + 1)/2h) \). \( \rho_j^w = 0 \) represent the situations of purely random classifications. \( \rho_j^w = 1 \) represent perfectly
repeatable classifications. $\rho^w$ is the average value of $\widetilde{\rho}_j^w$. The indicator $\pi^w_j$ of the repeatability is expressed as follows. $\pi^w_j(h)$ expresses the extent of matching between the measured value and the threshold value of the category within a laboratory $j$. $\pi^w_j$ is the value obtained as a probability that the true value is included within the range of the category boundaries. $\pi^w_j(h)$ is defined as eq (4).

$$\pi^w_j(h) = P(Y_{ijk} = h|\delta_{j,h} < X_i < \delta_{j,h+1}) = \frac{\int_{x=\delta_{j,h}}^{\delta_{j,h+1}} q_j(h|x)\varphi(x)dx}{\int_{x=\delta_{j,h}}^{\delta_{j,h+1}} \varphi(x)dx}$$  \hspace{1cm} (4)

$\widetilde{\pi}_j^w$ is a value obtained by rescaling $\pi^w_j$ to $[0, 1]$. It is calculated $\widetilde{\pi}_j^w = (\pi_j^w - 1/h)(1 - 1/h)$.

The indicator $\rho^b_{j_1,j_2}$ ($b$ means between) expresses whether the ordering of the measured values match between laboratories $j_1$ and $j_2$. $\rho^b_{j_1,j_2}$ is defined as eq (5).

$$\rho^b_{j_1,j_2} = P(Y_{ij_1} \leq Y_{ij_2}|X_i \leq X_u)$$  \hspace{1cm} (5)

$\rho^b$ is the average value of $\rho^b_{j_1,j_2}$. $\rho^b$ is calculated $\rho^b = (\rho^b - (h + 1)/2h)(1 - (h + 1)/2h)$. The indicator $\pi^b_{j_1,j_2}$ expresses the extent of matching between the thresholds of the categories $h$ and $h - 1$ and the measured values between laboratories. $\pi^b_{j_1,j_2}$ is defined as eq (4).

$$\pi^b_{j_1,j_2} = \sum_{h=1}^{H} P(\delta_{j_1,h-1} < X < \delta_{j_1,h} \wedge \delta_{j_2,h-1} < X < \delta_{j_2,h})$$

$$= \sum_{h=1}^{H} \max\{0, \Phi\{\min(\delta_{j_1,h}, \delta_{j_2,h})\} - \Phi\{\max(\delta_{j_1,h-1}, \delta_{j_2,h-1})\}\}$$  \hspace{1cm} (6)

$\pi^b$ is the average value of $\pi^b_{j_1,j_2}$. $\widetilde{\pi}^b$ is calculated $\widetilde{\pi}^b = (\pi^b - 1/h)(1 - 1/h)$. Note that $\varphi(x)$ and $\Phi(x)$ are the density and distribution functions of the standard normal distribution, respectively.

4. Simulation experiment of measurement accuracy in IRT

4.1 Analytical purpose and procedure

We investigated whether the IRT model applied to ISO 5725 and the measurement-accuracy indicators can evaluate the accuracy of the measurement of ordinal-categorical data. First, we configure different conditions and create measurement value data. Then, we estimate the parameters of GPCM, which apply the created data to ISO 5725, and calculate the indicators of measurement accuracy. Furthermore, from the calculation results, we examine whether the measurement accuracy of the ordinal-categorical data can be evaluated.

4.2 Condition setting

ISO 5725 pertains to collaborative-assessment experiments; however, in this simulation, it is assumed that certain laboratories measure certain samples. Samples are assumed to have different true values and follow the standard normal distribution. Each laboratory classifies samples into four categories on an order scale: [0: bad; 1: normal; 2: good; 3: best]. In classifying samples, consideration will be given to the variation in the bias (trueness), the variation within a laboratory (repeatability), and the variation between laboratories (reproducibility). The number of samples is set to 100, 250, 500 and the number of laboratories is set to 10, 25, 50. Simulations are performed 100 times under the above conditions.
4.3 Creation of sequence-categorical data

\( Y_{ij} \) is a measurement value obtained by measuring the sample \( i \) in laboratory \( j \); it is represented by

\[ Y_{ij} = \mu_i + k + B_j + e_{ij}. \] \( (7) \)

\( \mu_i \) is the true value of sample \( i \), \( k \) is bias, which is not dependent on the laboratory, \( B_j \) is variation due to laboratory \( j \), and \( e_{ij} \) is a random error associated with measuring the sample \( i \) in laboratory \( j \). We assume that \( \mu_i \) follows \( N(0,1^2) \), \( B_j \) follows \( N(0,\sigma_b^2) \), and \( e_{ij} \) follows \( N(0,\sigma_e^2) \).

For each laboratory, measurements are determined by randomly generated values according to the abovementioned distribution and classified into four categories. The measured value is classified as Category 0 for \( Y_{ij} \leq -1 \), Category 1 for \( -1 < Y_{ij} \leq 0 \), Category 2 for \( 0 < Y_{ij} \leq 1 \), and Category 3 for \( 1 < Y_{ij} \). The standard deviations, \( \sigma_w, \sigma_b \), represent variations in measured values. In this simulation, we set \( \sigma_w = 0, 0.5, 1.0 \) to create variation within a laboratory and evaluate repeatability. Furthermore, we set \( \sigma_b = 0, 0.25, 0.5, 1, 2 \) to create variation between laboratories and evaluate reproducibility.

There are various cases in the real problem about the magnitude relation of the variation between variation due to laboratory \( B_j \) and random error \( e_{ij} \). In experiments decided in detail, variation between laboratories is smaller than variation within the laboratory. On the other hand, in the case where the standardization of the test procedure is not sufficient or in the case of using an organism such as an animal, the variation between laboratories is larger than the variation within the laboratory. For example, the variation within the laboratory becomes large because the pathological data of animal experiments have individual differences among creature. Furthermore, the variation between laboratories may be more than twice as large as the variation within the laboratory the reasons such as not being able to arrange the same equipment and environment. Therefore, \( \sigma_w = 1.0 \) and \( \sigma_b = 2.0 \) were also set. \( \sigma_w = 0 \) and \( \sigma_b = 0 \) assume that the classification within a laboratory and the classification between laboratories are perfectly consistent. We also set \( k = 0, 0.5, 1.0 \) to create measurement bias and evaluate the trueness.

First, Table 1 shows the categories and true values, which are determined by the standard normal random number.

Table 1. Categories classified according to the true values for each sample
(The number of samples is set to 100.)

| sample i | true value_i | category_i |
|----------|--------------|------------|
| 1        | 1.93         | 3          |
| 2        | -0.20        | 1          |
| 3        | -0.82        | 1          |
| 4        | 0.02         | 2          |
| 5        | 0.83         | 2          |
| :        | :            | :          |
| 100      | 1.71         | 3          |

Table 1 shows the results of the true values \( \mu_i \) of sample \( i = 0, 1, 2, \cdots, 100 \), as determined by standard normal random numbers and categories classified into four scales, \( h = 0, 1, 2, 3 \).

Next, we generate random numbers following \( N(0,\sigma_w^2) \), \( N(0,\sigma_b^2) \) and add them to the determined true values of \( \mu_i \) and bias to constitute the measured values, \( Y_{ij} \). Table 2 shows some of the classification results for the measured values, \( Y_{ij} \).

Table 2. Classification results for \( \sigma_w = 0, \sigma_b = 0.5, k = 0 \)
(The number of samples is set to 100 and the number of laboratories is set to 10.)
Table 2 shows the results for the classification of the measurement values $Y_{ij}$ into categories in each laboratory as $\sigma_w = 0$, $\sigma_b = 0.5$, and $k = 0$.

### 4.4 Results and consideration of changing the variation

First, we consider parameters estimation, the categorical-probability curves, and measurement accuracy when variations are changed. The number of samples is set to 100 and the number of laboratories is set to 10. For the data created in Table 2, the item-discrimination parameter, $\alpha_j$, and the item-threshold parameter, $\delta_{jm}$, for each laboratory were estimated using the R package. The average values of the results of the estimated parameters for each laboratory are shown in Tables 3, 4, and 5.

| sample i | laboratory1 | laboratory2 | laboratory3 | ... | laboratory10 |
|----------|-------------|-------------|-------------|-----|--------------|
| 1        | 3           | 2           | 3           | ... | 3            |
| 2        | 1           | 2           | 1           | ... | 0            |
| 3        | 0           | 1           | 1           | ... | 1            |
| 4        | 2           | 1           | 2           | ... | 1            |
| 5        | 3           | 2           | 3           | ... | 2            |
| ...      | :           | :           | :           | :   | :            |
| 100      | 3           | 3           | 3           | ... | 3            |

As shown in Tables 3, 4, and 5, since the item-threshold parameter $\delta_{jm}$ is approximately $\delta_{j,1} < \delta_{j,2} < \delta_{j,3}$, this indicates that the category increases along with the true value. However, as the variance increases, the values of $\delta_{j,1}, \delta_{j,2}, \delta_{j,3}$ are sometimes reversed.

The smaller the value of the item-discrimination parameter $\alpha_j$, the larger the variation. This indicates that the measurement accuracy decreases when the ability to classify the measured values into categories is low. Moreover, when the value of $k$ changes, the values of parameter did not change very much. Therefore, it is believed that the bias does not affect classification ability.

[DOI : 10.17929/tqs.4.22]
The parameters estimated in the previous section were assigned to GPCM and a categorical-probability curve was created for each laboratory. The Category-probability curve of 10 laboratories with each variability pattern are shown in Figures 3, 4, and 5. Based on the values of $\sigma_w$, $\sigma_p$ was changed and the characteristics were observed.

| $\sigma_p$ | $\sigma_p$ | $\sigma_p$ | $\sigma_p$ |
|-----------|-----------|-----------|-----------|
| $\sigma_w=0$ | $\sigma_w=0$ | $\sigma_w=0$ | $\sigma_w=0$ |
| $\sigma_p=0.25$ | $\sigma_p=0$ | $\sigma_p=1.0$ | $\sigma_p=2.0$ |

Figure 3. Category-probability curve ($\sigma_w = 0$)

Figure 4. Category-probability curve ($\sigma_w = 0.5$)

Figure 5. Categorical-probability curve ($\sigma_w = 1.0$)
It can be seen from Figures 3, 4, and 5 that the larger the variation, the smaller the slope of the tangent line of $\theta = \delta_{jm}$. Moreover, as the bias increases, the curve moves to the left. By observing the categorical-probability curve for each laboratory, the results of parameters $\delta_{jm}$, $\alpha_j$ were visually confirmed.

Tables 6, 7, and 8 show the calculated results of the measurement-accuracy indicators of the correct-ordering probability, $\rho$, and the correct classification probability, $\pi$. These are both repeatable and reproducible.

### Table 6. Measurement-accuracy indicators ($k = 0$)

| $k = 0.0$ | $\sigma_w=0$ | $\sigma_w=0.25$ | $\sigma_w=0.5$ | $\sigma_w=1.0$ | $\sigma_w=1.0$ | $\sigma_w=2.0$ | $\sigma_w=1.0$ | $\sigma_w=2.0$ | $\sigma_w=1.0$ | $\sigma_w=2.0$ | $\sigma_w=1.0$ | $\sigma_w=2.0$ |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| $\rho_w$  | 0.82      | 0.69      | 0.50      | 0.33      | 0.69      | 0.66      | 0.60      | 0.47      | 0.33      | 0.50      | 0.48      | 0.47      |
| $\rho_b$  | 0.82      | 0.68      | 0.49      | 0.31      | 0.68      | 0.65      | 0.59      | 0.45      | 0.31      | 0.49      | 0.47      | 0.46      |
| $\pi_w$   | 0.76      | 0.57      | 0.36      | 0.30      | 0.57      | 0.53      | 0.46      | 0.33      | 0.30      | 0.36      | 0.35      | 0.34      |
| $\pi_b$   | 0.90      | 0.84      | 0.64      | 1.41      | 0.84      | 0.82      | 0.77      | 0.63      | 1.39      | 0.65      | 0.64      | 0.63      |

### Table 7. Measurement-accuracy indicators ($k = 0.5$)

| $k = 0.5$ | $\sigma_w=0$ | $\sigma_w=0.25$ | $\sigma_w=0.5$ | $\sigma_w=1.0$ | $\sigma_w=2.0$ | $\sigma_w=1.0$ | $\sigma_w=2.0$ | $\sigma_w=1.0$ | $\sigma_w=2.0$ | $\sigma_w=1.0$ | $\sigma_w=2.0$ |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| $\rho_w$  | 0.82      | 0.69      | 0.51      | 0.33      | 0.69      | 0.66      | 0.61      | 0.48      | 0.33      | 0.51      | 0.49      |
| $\rho_b$  | 0.82      | 0.68      | 0.35      | 0.32      | 0.68      | 0.66      | 0.60      | 0.47      | 0.32      | 0.50      | 0.47      |
| $\pi_w$   | 0.76      | 0.59      | 0.38      | 0.31      | 0.59      | 0.55      | 0.48      | 0.36      | 0.31      | 0.38      | 0.35      |
| $\pi_b$   | 0.89      | 0.85      | 0.66      | 1.41      | 0.84      | 0.82      | 0.77      | 0.63      | 1.39      | 0.65      | 0.64      |

### Table 8. Measurement-accuracy indicators ($k = 1.0$)

| $k = 1.0$ | $\sigma_w=0$ | $\sigma_w=0.25$ | $\sigma_w=0.5$ | $\sigma_w=1.0$ | $\sigma_w=1.0$ | $\sigma_w=2.0$ | $\sigma_w=1.0$ | $\sigma_w=2.0$ | $\sigma_w=1.0$ | $\sigma_w=2.0$ | $\sigma_w=1.0$ | $\sigma_w=2.0$ |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| $\rho_w$  | 0.81      | 0.71      | 0.53      | 0.34      | 0.71      | 0.69      | 0.63      | 0.50      | 0.32      | 0.52      | 0.50      |
| $\rho_b$  | 0.81      | 0.70      | 0.52      | 0.31      | 0.70      | 0.68      | 0.62      | 0.49      | 0.31      | 0.52      | 0.48      |
| $\pi_w$   | 0.76      | 0.65      | 0.43      | 0.30      | 0.65      | 0.62      | 0.53      | 0.41      | 0.30      | 0.43      | 0.41      |
| $\pi_b$   | 0.90      | 0.87      | 0.71      | 1.64      | 0.87      | 0.86      | 0.80      | 0.71      | 1.42      | 0.71      | 0.64      |

According to Tables 6, 7, and 8, there was no particular change in the value due to the bias. Regarding variation, the smaller the indicator values in most patterns, the larger the variation. Therefore, lower accuracy is considered to lead to smaller indicator values. However, was larger than the pattern with small variance at $\sigma_w = 1.0$, $\sigma_b = 0.5$ and $\sigma_w = 1.0$, $\sigma_b = 1.0$. The value of $\pi_b$ exceeded 1 at $\sigma_w = 0, 0.5, 1.0, \sigma_b = 2.0$. $\pi_b$ takes a value in the range $0 \leq \pi_b \leq 1$, and the closer it is to 1, the more accurate it is. However, the value of $\pi_b$ becomes larger even if it exceeds 1 or the variation is large.

From equation (6), $\pi_{b,h}$ is the common area between the thresholds of any two laboratory values ($\delta_{j,h}$ and $\delta_{j,h+1}$) and the standard normal distribution. The filled area in Figure 6 is a common area, and if the entire area of the distribution is common, the common area is 1. However, when the variation becomes larger, as shown in Figure 7, the threshold may be reversed or the common area may overlap among categories. This causes the value of $\pi_{b,h}$ to exceed 1 or to increase despite large variations. Therefore, it was found that the indicator $\pi_{b,h}$ had problems evaluating the reproducibility of the measurement results.
In the real problem, from the category probability curve of each laboratory, it is possible to visually grasp the variation within the laboratories and the variation between the laboratories by slope and threshold value. With $\sigma_b < 1.0$, we can get the degree of agreement of order and the degree of matching of thresholds from $\rho$ and $\pi$. If the variation between laboratories is $\sigma_b > 1.0$, there is a possibility that the thresholds are interchanged. In this IRT model, it is a problem that standard deviation as assumed in simulation cannot be obtained.

4.5 Results and consideration of changing the sample size

We consider variations in measurement accuracy when changing the number of samples. In this simulation, we set $\sigma_w = 0.25$ to create variation within a laboratory and $\sigma_b = c\sigma_w$ ($c = 0.25$) to create variation between laboratories. The number of samples considered are 100, 250, and 500. Figure 8 is a histogram of values obtained by estimating the measurement-accuracy indicators 100 times for each number of samples. The larger the number of samples is, the smaller the dispersion of the values becomes. Therefore, as the number of samples increases, the measurement accuracy can be better estimated.

$\pi_b$ approaches 1 as the number of samples increases. As the number of samples becomes larger, the smaller the variation among the laboratory thresholds $\delta_{jm}$ becomes as a characteristic of IRT. Therefore, $\pi_b$ is considered to improve the measurement accuracy as the number of samples increases.
5. Conclusion

In this study, we examined whether IRT is suitable as a method for evaluating the accuracy of measurement of ordinal-categorical data. Different conditions were set and analyzed by simulation. It was found to be possible to evaluate variation of precision using the correct-ordering probability $\rho$ and correct classification probability $\pi$, which are the indicators of IRT. However, the reproducibility of $\pi$ was found to be problematic for accuracy evaluation as the variation increased. In addition, it was possible to find the change in trueness from the parameters $\delta_{jm}, \alpha_j$ and the categorical-probability curves. However, it is not yet possible to evaluate the trueness with a specific value. In addition, we examined the characteristics of measurement-accuracy indicators considering sample number.

As future research, we need to improve the equation for $\pi$'s reproducibility, compare it with the various methods proposed in the previous research, and consider a method suitable for evaluating measurement accuracy. In addition, we believe it is necessary to consider indicators for evaluating trueness. Also, the larger the number of samples, the more stable the estimate of the parameter, 500 was the best in the simulation. However, it may be difficult to secure the number of samples in a real problem. It is necessary to consider how to determine the number of samples in the future.

Reference

Blair J., Lacy M. G. (2000), “Statistics of ordinal variation”, Sociological Methods & Research, Vol. 28, pp. 251–280
Bashkansky E., Gadrich T., Kuselman I. (2012), “Interlaboratory comparison of test results of an ordinal or nominal binary: analysis of variation”, Accreditation and Quality Assurance, Vol. 17, pp. 239-444
ISO 5725-1 (1994), Accuracy (trueness and precision) of measurement methods and results Part 1: General principles and definitions
ISO 5725-2 (1994), Accuracy (trueness and precision) of measurement methods and results Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method
De Mast J., Van Wieringen W. N. (2010), “Modeling and Evaluating Repeatability and Reproducibility of Ordinal
Evaluation of the accuracy, M. Ogawa et al.

Classifications,” Technometrics, Vol. 52, No. 1, pp. 94–106
Muraki E. (1992), “A Generalized Partial Credit Model: Application of an EM Algorithm,” Applied Psychological Measurement, Vol. 16, pp. 159–176
Suzuki T., Katakura A., Miyazawa R., Sano N. (2015), “Precision Evaluation for Non-quantitative Measurements—Binary and Ordinal Categorical Cases—,” Proc. XXI IMEKO World Congress
Masters G. N. (1982), “A Rasch Model for Partial Credit Scoring,” Psychometrika, Vol. 47, pp.149–174.
Samejima F. (1969), “Estimation of Latent Ability Using a Response Pattern of Graded Scores,” Psychometrika, Vol. 38, pp. 203-219
Lord F. (1980), Applications of Item Response Theory to Practical Testing Problems, Hillsdale: Lawrence Erlbaum.

Acknowledgements:
The authors would like to thank anonymous referees for their suggestions comments.

Author’s biographical notes

Mayu ogawa is a graduate student at Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science. Her research has focused on statistical data analysis and statistical quality control.

Mirai Tanaka is an assistant professor at Department of Mathematical analysis and Statistical Inference, The Institute of Statistical Mathematics. His research has focused on operations research and data analysis.

Xiao-Nan Lu is an assistant professor at Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science. His research has focused on Discrete Mathematics (Combinatorics) and related interdisciplinary research emphasis on Statistics and Informatics.

Tomomichi Suzuki is a professor at Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science. His research has focused on statistical data analysis and statistical quality management.

[DOI : 10.17929/ tqs.4.22]

Received, April 7, 2017
Revised: May 28, 2018
Accepted: June 8, 2018