Proceedings of the 15th Annual Conference of the Japan Association for Research on Testing, 2017, 152－155 (manuscript)

Three Aspects of Reliability in Item Response Theory

Visual representations of reliability

○Yasuharu Okamoto

Faculty of Integrated Arts and Social Sciences, Japan Women's University

1. Introduction

Reliability has mainly been discussed in classical test theory (CTT), and coefficients have been defined according to three aspects, precision, stability, and predictive power. Precision comes from the ratio of the variance of true scores to that of observed scores and is considered the mathematical definition of reliability. Stability is denoted by correlation coefficient between two parallel tests. Predictive power is represented by strength of relation between observed values and true values and is obtained by the regression model's coefficient of determination for observed values and true values as dependent and predictor variables, respectively. These three coefficients of reliability coincide in the case of continuous items. However, in the case of categorical items, they differ. Results of simulation (Okamoto, 2016) showed that alpha coefficient tends to be larger than strength of relation, i.e., coefficient of determination, between true values and observed values in the case of categorical items. These results indicate the necessity of considering reliability in item response theory (IRT) from the three points of view.

2. Three Coefficients of Reliability in IRT

Explicit definitions of coefficients of reliability need mathematical models for IRT. This study adopts the following notation for mathematical modeling.

X_i: Test item i, which may have more than two response categories.

$\mathrm{S}(X_{p(i)}) = \{X_{p(1)}, \cdots, X_{p(m)}\}$: Used items for person p, i.e., $X_{p(j)}$ denotes the j-th item from the test items X_is, which person p takes. In the case of a fixed set of test items, $\mathrm{S}(X_{p(i)})$ is the constant set of test items, i.e., the set of all test items. In the case of an adaptive test (e.g., Okamoto, 2007), $\mathrm{S}(X_{p(i)})$ depends on the performance of person p, so is considered to be determined stochastically depending on person p's performance.

θ_p: The true value of person p's ability, which may be represented by a vector when a multifactor model is employed. In this study, a one-factor model is considered, because it is sufficient to use a simple model to show this study's essential point.

$\hat{\theta}_p$: Estimate of θ_p. Popular methods of estimation are the maximum likelihood and the Bayesian. When the person's responses are all the same extreme ones, $\hat{\theta}_p$ cannot be estimated by the maximum likelihood method without ad hoc assumptions. Hence, this study

employs the Bayesian method.

The modelling above is represented by the following diagram.

$$\theta_p \quad \rightarrow \quad \mathrm{S}\left(X_{p(i)}\right) = \left\{X_{p(1)}, \cdots, X_{p(m)}\right\} \quad \rightarrow \quad \hat{\theta}_p \tag{1}$$

Using diagram (1), three types of coefficients of reliability, precision, stability, and predictive power, are given as follows.

2.1. Precision. Precision is the popular mathematical definition of coefficient of reliability. In IRT, this coefficient is given for the maximum likelihood (ML) method (Toyoda, 1989). For the Bayesian method, the coefficient of reliability is proposed as given by

$$\rho_{V(\theta_p)} = 1 - V_{post(\theta_p)} / V_{prior(\theta_p)} \tag{2}$$

where $V_{post(\theta_p)}$ and $V_{prior(\theta_p)}$ are variances of post and prior distributions of θ_p (Okamoto, 2015). $V_{prior(\theta_p)}$ and $V_{post(\theta_p)}$ correspond to variances of observed scores and error terms in the case of CTT, respectively.

Taking expectation of $\rho_{V(\theta_p)}$, we have

$$\bar{\rho}_{V(\theta)} = E\left(\rho_{V(\theta_p)}\right) \tag{3}$$

2.2. Stability. Stability can be represented by correlation coefficient between two parallel tests. A sequence of a parallel test for sequence (1) can be represented by diagram (1')

$$\theta_p \quad \rightarrow \quad \mathrm{S}\left(X'_{p(i)}\right) = \left\{X'_{p(1)}, \cdots, X'_{p(m)}\right\} \quad \rightarrow \quad \widehat{\theta'}_p \tag{1'}$$

Person p takes two parallel tests, and results are denoted by $\mathrm{S}\left(X_{p(i)}\right)$ and $\mathrm{S}\left(X'_{p(i)}\right)$, respectively. Corresponding to $\mathrm{S}\left(X_{p(i)}\right)$ and $\mathrm{S}\left(X'_{p(i)}\right)$, two estimates $\hat{\theta}_p$ and $\widehat{\theta'}_p$ are given. Test stability is given by the correlation coefficient between $\hat{\theta}_p$ and $\widehat{\theta'}_p$.

$$\rho_{\hat{\theta}_p \hat{\theta}'_p} = Cor\left(\hat{\theta}_p, \hat{\theta}'_p\right) \tag{4}$$

Samejima (1994) discusses correlation (4) in the framework of the ML method.

2.3. Predictive Power. Prediction of the true value θ_p from the estimate $\hat{\theta}_p$ can be represented as regression model

$$\theta_p = a_1 \hat{\theta}_p + b_1 \tag{5}$$

The predictive power of model (5) is given by the coefficient of determination R^2 of model (5), which equals the square of the correlation coefficient between θ_p and $\hat{\theta}_p$.

$$R^2_{\theta_p \hat{\theta}_p} = \left\{Cor\left(\theta_p, \hat{\theta}_p\right)\right\}^2. \tag{6}$$

Model (5) represents how well the true value θ_p can be inferred from the estimate $\hat{\theta}_p$. On the other hand, the following model (7) represents how strongly the estimate $\hat{\theta}_p$ is constrained by the true value θ_p.

$$\hat{\theta}_p = a_2 \theta_p + b_2 \tag{7}$$

The coefficient of determination of model (7) is the same as that of model (5). The coefficient $R^2_{\theta_p \hat{\theta}_p}$ denotes strength of relation between the true value θ_p and the estimate $\hat{\theta}_p$, which is a common value for models (5) and (7).

3. Examples by Simulation

Proceedings of the 15th Annual Conference of the Japan Association for Research on Testing, 2017, 152－155 (manuscript)

Some examples of $\rho_{V(\theta_p)}$, $\rho_{\hat{\theta}_p\hat{\theta}'_p}$, and $R^2_{\theta_p\hat{\theta}_p}$ are presented below.

Consider M test items X_is, $i = 1, \cdots, M$, which are all binary, i.e., $X_i = 0$ or 1. For them, set the two-parameter logistic model

$$P(X_i = 1|\theta) = 1/\{1 + exp(-1.7a_i(\theta - b_i))\}. \quad (8)$$

In the following simulation, the fixed set of test items $X_1, \cdots, X_M$ is used for all persons under each condition, i.e., the test is not adaptive.

$$S(X_{p(i)}) = \{X_1, \cdots, X_M\}$$

Parameter values of a_is and b_is are set as

$$a_i = 1 \qquad and \qquad b_i = 0$$

Prior distribution for θ_p is assumed to be the standard normal distribution $N(0, 1)$.

With these settings, two simulations for precision $\sigma_{V(\theta_p)}$ and one simulation for each stability $\rho_{\hat{\theta}_p\hat{\theta}'_p}$ and predictive power $R^2_{\theta_p\hat{\theta}_p}$ were run. In this study, a point estimate of $\hat{\theta}_p$ (or $\hat{\theta}'_p$) was given by the posterior distribution's median.

3.1 Precision. For each number of test items $M = 5$ or 100, seven values of $\theta_p = -3, -2, -2, 0, 1, 2, 3$ were used. For each value of θ_p, 20 experiments were completed. Each experiment consisted of application of items X_is to a person with θ_p, estimation of a posterior distribution, and calculation of precision $\rho_{V(\theta_p)}$. Scattergrams of points $\left(\theta_p, \rho_{V(\theta_p)}\right)$ for $M = 5$ and 100 are shown in Figures 1 and 2, respectively. Small random fluctuations were added horizontally to positons $\left(\theta_p, \rho_{V(\theta_p)}\right)$, so that multiple points with the same position do not completely overlap.

3.2 Stability. For each value of θ_p, which was sampled from the standard normal distribution, two esimates $\hat{\theta}_p$ and $\hat{\theta}'_p$ were calculated for $M = 5$ items according to parallel diagrams (1) and (1'),

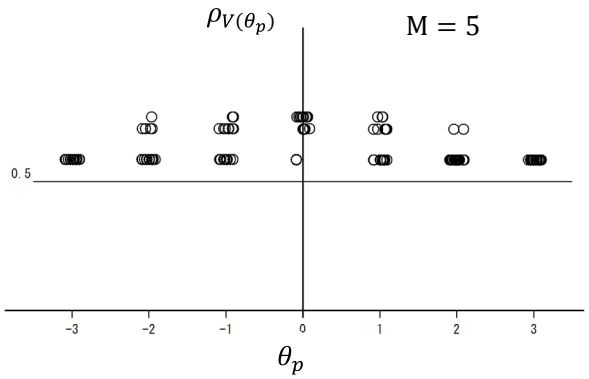

Figure 1. Scattergram of points $\left(\theta_p, \rho_{V(\theta_p)}\right)$ for M = 5.

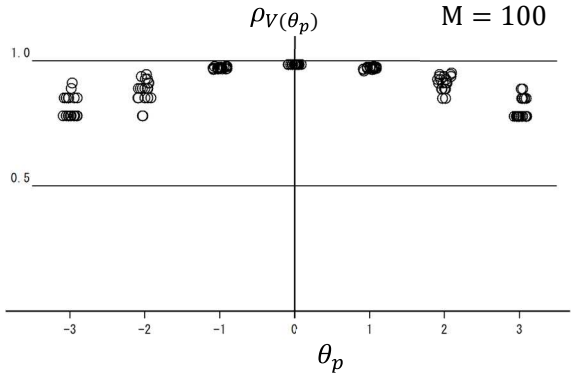

Figure 2. Scattergram of points $\left(\theta_p, \rho_{V(\theta_p)}\right)$ for M = 100.

respectively. A scattergram of $(\hat{\theta}_p, \hat{\theta}'_p)$s with small random perturbations for 1000 samples θ_p s is shown in Figure 3. The reliability coefficient for stability is $\rho_{\hat{\theta}_p\hat{\theta}'_p} = r = 0.720$.

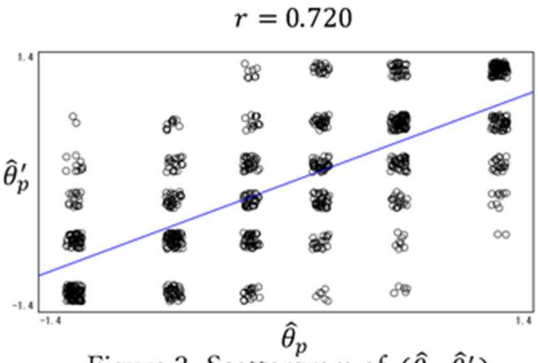

Figure 3. Scattergram of $(\hat{\theta}_p, \hat{\theta}'_p)$

3.3 Predictive Power. For each value of θ_p, sampled from the standard normal distribution, an esimate $\hat{\theta}_p$ was calculated for $M = 5$ items. A scattergram of $(\hat{\theta}_p, \theta)$ s for 1000 samples θ_ps is shown in Figure 4. The reliability coefficient for predictive power is $R^2_{\theta_p\hat{\theta}_p} = r^2 = 0.682$.

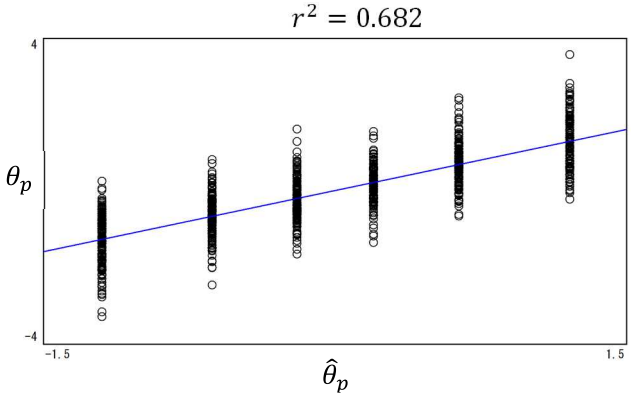

Figure 4. Scattergram of $(\hat{\theta}_p, \theta_p)$.

4. Conclusion

Visual displays such as scattergrams in Figures 1 to 4 demonstrate information on reliability, which can help us intuitively evaluate estimates of ability.

5. References

Okamoto, Y. (2007). エントロピー最小化基準による適応型テストーテスト情報関数の問題点ー [Adaptive test by minimum entropy criterion: A problems of test information function]. 日本テスト学会誌[*Japanese Journal for Research on Testing*], 3, 36-47.

Okamoto, Y. (2015). IRT における信頼性係数について[On a reliability coefficient in IRT].日本テスト学会第 13 回大会発表論文抄録集[*Proceedings of the 13th Annual Conference of The Japan Association for Research on Testing*], 72-75.

Okamoto, Y. (2016). Reliability Coefficients for Ordinal Categorical Items: Actual relation of observed and true scores. *Japan Women's University Journal: Faculty of Integrated Arts and Social Sciences,* 2016, 27, 113-122.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement*, 18, 229-244.

Toyoda, H. (1989). 項目反応理論における信頼性係数の推定法[The methods for estimating the reliability coefficient under item response model]. 教育心理学研究[*Japanese Journal of Educational Psychology*], 37, 283-285.