Reliability Coefficients for Ordinal Categorical Items: Actual Relation of Observed and True Scores

順序カテゴリ項目の信頼性係数　―観測値と真値の真の関係―

OKAMOTO Yasuharu
岡　本　安　晴

[Abstract] The coefficient for ordinal categorical items that represents the strength of the relation between observed scores and true values is proposed and compared with reliability coefficients by simulation. For continuous items, the relational strength of observed scores and true values can be represented by reliability coefficients. However, for ordinal categorical items, an observed categorical response is modeled according to Thurstone: This type of model requires a coefficient for ordinal categorical items to be provided differently from that for continuous items. The proposed coefficient for ordinal categorical items is a coefficient of determination for a regression model, which represents the relation between observed scores and true values. Results of simulations of coefficients for ordinal categorical items indicate that reliability coefficients, the correlation coefficient of parallel tests, and the coefficient alpha overestimate actual relationships of observed scores and true values, which the proposed coefficient adequately represents.

[Key Words] reliability coefficient, ordinal categorical items, coefficient of determination, regression model

Introduction

The coefficient that represents the actual relationship between a true value and an observed score given as a sum of response categories on test items is proposed: it is shown by simulation that popular reliability coefficients, coefficient alpha and correlation coefficient of parallel tests, overestimate the actual relationship in cases of ordinal categorical items. The term *reliability* is used to refer to consistency through a series of measurements (Cronbach, 1961) or to the precision with which the test score measures the attribute (McDonald, 1999). A test score or measurement X is assumed to be represented as a sum of a true value T and an error E_{rr}, which is independent of T. That is, we have

$$X = T + E_{rr}. \tag{1}$$

Consistency can be represented by the correlation coefficient $\rho_{XX'}$ of scores X and X' on parallel tests (Guttman, 1945), i.e.,

$$\rho_{XX'} = Cor(X, X'),$$

where $Cor(X, X')$ denotes a correlation coefficient of X and X'.

Precision ρ of test score X is represented by the ratio of variance of a true score T to variance of a test score X, i.e.,

$$\rho = \frac{Var(T)}{Var(X)}.$$

$Var(X)$ and $Var(T)$ denote variances of X and T, respectively. As we know, we have

$$\rho = \rho_{XX'}.$$

When X is a sum of scores X_js on items j; $j = 1, \cdots, M$, the coefficient alpha α is given by

$$\alpha = \frac{M}{M-1}\left\{1 - \frac{\sum_{j=1}^{M} Var(X_j)}{Var(X)}\right\}.$$

Now, assume that X_j is represented by a single factor model

$$X_j = \mu_j + \lambda_j F + E_j, \tag{2}$$

where F is a common factor with the standard normal distribution, λ_j is a factor loading, and E_j is an error with a normal distribution with mean 0 and variance ψ_j^2. Error terms E_js are independent of each other and F. Parameters μ_j and λ_j correspond to location and discrimination parameters in item response theory (IRT). In case of Model 2, total score X is given by

$$X = \sum_{j=1}^{M} X_j = \sum_{j=1}^{M}(\mu_j + \lambda_j F) + \sum_{j=1}^{M} E_j = T + E_{rr}\,, \tag{3}$$

where $T = \sum_{j=1}^{M}(\mu_j + \lambda_j F)$ is a true score and $E_{rr} = \sum_{j=1}^{M} E_j$ is an error.

For Model 3, reliability coefficient ρ is given by coefficient omega ω (McDonald, 1999)

$$\omega = \frac{Var(T)}{Var(X)} = \frac{\left(\sum_{j=1}^{M}\lambda_j\right)^2}{\left(\sum_{j=1}^{M}\lambda_j\right)^2 + \sum_{j=1}^{M}\psi_j^2}. \tag{4}$$

By definition 4, we have

$$\rho = \omega.$$

In case of Model 2, as we know, when all λ_js have the same value, we have

$$\alpha = \omega.$$

The purpose of measurement is to measure an attribute. Estimation of a true value T of the attribute by an observed score X can be represented by the following regression model

$$T = a + bX + e, \tag{5}$$

where values of coefficient a and b are set such that expectation of squared error e^2 is minimized. In statistics, the precision of a regression model is indicated by the coefficient of determination R^2. For Model 5, R^2 is given by

$$R^2 = \frac{Var(a + bX)}{Var(T)}.$$

That is, R^2 represents the ratio of the variance of estimation $a + bX$ by the independent variable X to the variance of the dependent variable T in Model 5.

As is known, R^2 is equal to the squared correlation coefficient of T and X, i.e.,

$$R^2 = \{Cor(T, X)\}^2,$$

and we also have

$$\{Cor(T, X)\}^2 = \rho.$$

Hence, we have

$$R^2 = \rho = \omega = \rho_{XX'}. \qquad (6)$$

However, in the case of ordinal categorical items, Equation 6 does not hold.

Coefficients for Ordinal Categorical Items

In the case of continuous items, an observed value X is given as a sum of a true value T and an error E_{rr} (Model 1). In the case of ordinal categorical items, we place the model for ordinal categorical items in this paper as follows.

Let U_j be an unobserved continuous value of item j. A single factor model for U_j, the same as that for X_j (Equation 2), is set. That is,

$$U_j = \mu_j + \lambda_j F + E_j, \qquad (7)$$

where F is a common factor with the standard normal distribution, λ_j is a factor loading, μ_j is a position parameter, and E_j is an error that has normal distribution with mean 0 and variance ψ_j^2. Error terms E_js are independent of each other and F. The observed response Y_j on item j, based on U_j, is made according to the rule

$$Y_j = k, \qquad \text{if } C_{k-1} \leq U_j < C_k, \qquad (8)$$

where C_ks are category boundaries for the response Y_j, and

$$-\infty = C_0 < C_1 < \cdots < C_{K-1} < C_K = +\infty\,.$$

An observed score Y can be obtained as the sum of Y_js, that is,

$$Y = \sum_{j=1}^{M} Y_j\,. \qquad (9)$$

In the case of ordinal categorical items, an observed continuous variable X_j in Model 2 becomes unobserved and is denoted by U_j. Unobserved variable U_j is categorized to Y_j, which is observed.

Under the model shown above for ordinal categorical items, coefficients discussed in the previous section are given as follows.

Coefficient ω (Equation 4) is calculated for latent variables U_js on items as ρ_ω (Okamoto, 2013), i.e.,

$$\rho_\omega = \frac{\left(\sum_{j=1}^{M} \lambda_j\right)^2}{\left(\sum_{j=1}^{M} \lambda_j\right)^2 + \sum_{j=1}^{M} \psi_j^2}. \qquad (10)$$

The coefficient of correlation of parallel tests is proposed by Green and Yang (2009) as a reliability coefficient, denoted by $\rho_{YY'}$ in this paper. That is,

$$\rho_{YY'} = Cor(Y, Y'),$$

where Y' and Y are parallel tests.

Coefficient α for ordinal categorical items can be given as follows:

$$\alpha = \frac{M}{M-1}\left\{1 - \frac{\sum_{j=1}^{M} Var(Y_j)}{Var(Y)}\right\}.$$

We know that the coefficient α for binary items, i.e., $K = 2$, is identical to the reliability coefficient called Kuder Richardson 20 or KR 20 (Kuder & Richardson, 1937; cf. Crocker & Algina, 1986).

In the case of ordinal categorical items, estimation of a value F of the attribute by an observed score Y is represented by the following regression model:

$$F = \gamma + \beta Y + \varepsilon. \qquad (11)$$

The values of coefficients γ and β are determined such that expectation of squared error ε^2 is minimized. As is known, the coefficient of determination of Model 11 is given by

$$R^2 = \frac{Var(\gamma + \beta Y)}{Var(F)} = \{Cor(Y, F)\}^2.$$

Comparisons of ρ_ω, $\rho_{YY'}$, α, and R^2 by simulations are presented in the next section.

Comparisons by Simulations

Comparisons of reliability coefficients ρ_ω, $\rho_{YY'}$, α, and R^2 were conducted by simulations.

Values of μ_js were sampled independently from a uniform distribution $U[-1, 1]$, where $U[a, b]$ denotes a uniform distribution over an interval $[a, b]$. Factor loadings λ_j were sampled independently from $U[0.5, 0.95]$. Category boundaries C_ks were also sampled independently from uniform distributions, which were chosen corresponding to the number of categories K, selected to be 2 and 6. For $K = 2$, only one category boundary C_1 was needed and was sampled from $U[-0.5, 0.5]$. For $K = 6$, each value of five category boundaries, $C_1 \cdots C_5$, was sampled independently from the respective uniform distribution, i.e.,

$$C_k \sim U[-1.6 + 0.5k, -1.4 + 0.5k], \qquad k = 1, \cdots, 5.$$

At each trial of simulation, parameter values μ_js, λ_js, and C_ks were randomly sampled from respective distributions: then for this set of parameter values $\{\mu_j\text{s},\ \lambda_j\text{s},\ C_k\text{s}\}$, coefficients ρ_ω, $\rho_{YY'}$, α, and R^2 were calculated and one set of coefficient values $\{\rho_\omega,\ \rho_{YY'},\ \alpha,\ R^2\}$ was made.

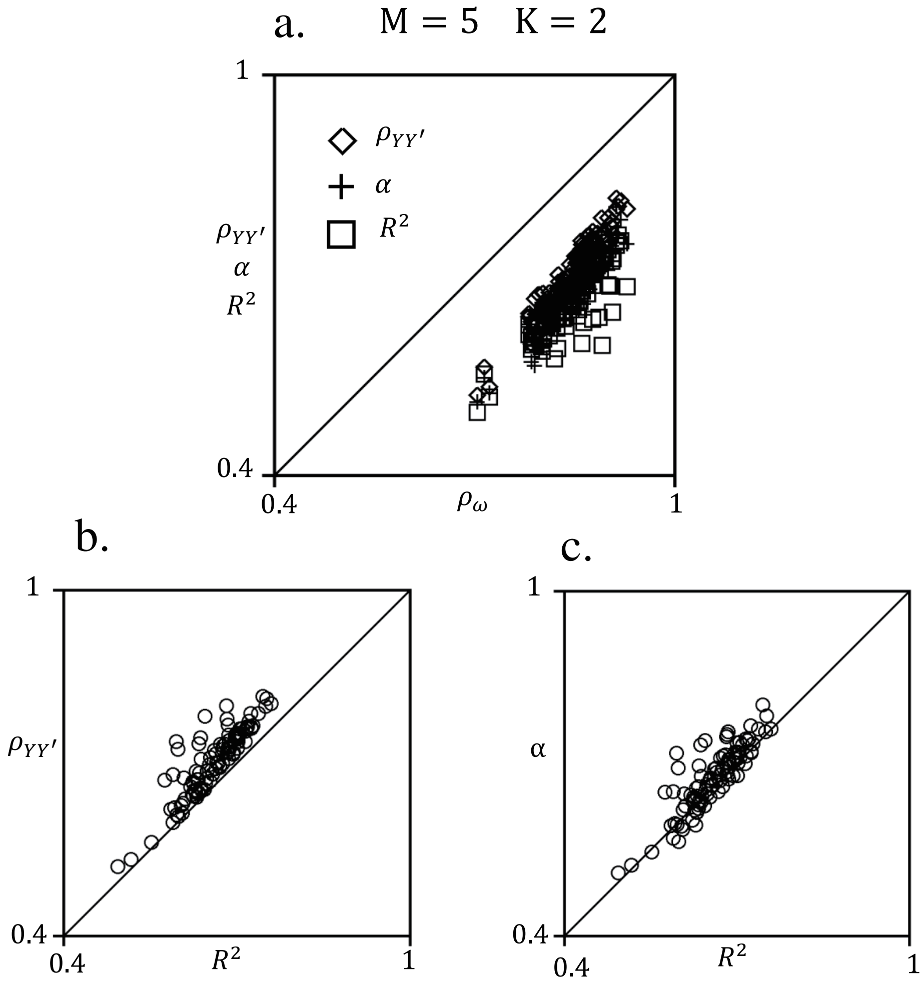

Figure 1. Scatter diagrams of points generated by 100 simulations for the number of items M = 5 and the number of categories K = 2. In each simulation, parameters were sampled randomly with the following distributions: $\mu_j \sim U[-1, 1]$, $\lambda_j \sim U[0.5, 0.95]$, and $C_1 \sim U[-0.5, 0.5]$, where $U[a, b]$ denotes a uniform distribution on an interval $[a, b]$. Figure 1a displays 100 sets of 3 points $(\rho_\omega, \rho_{YY'})$, (ρ_ω, α), and (ρ_ω, R^2). Figure 1b displays 100 points of $(R^2, \rho_{YY'})$. Figure 1c displays 100 points of (R^2, α).

Figure 1 shows 100 sets of coefficients. Parameter values μ_js, λ_js, and C_ks were sampled independently for each of 100 sets under the condition of the number of items $M = 5$ and the number of categories $K = 2$. Figure 1a shows a scatter diagram of 100 sets of points; each set consists of three points, $(\rho_\omega, \rho_{YY'})$, (ρ_ω, α), and (ρ_ω, R^2) corresponding to each combination of parameter values μ_j, λ_j, and C_k. All points are below the diagonal line, that is, $\rho_{YY'}$s, αs, and R^2s are smaller than corresponding ρ_ωs. These differences of $\rho_{YY'}$s, αs, and R^2s from ρ_ωs reflect loss of information by categorization Y_js of continuous variables U_js. Figure 1b shows a scatter diagram of 100 points of $(R^2, \rho_{YY'})$, all of which are above the diagonal line. Hence, $\rho_{YY'}$ overestimates strengths of actual relations of true values and observed scores. Figure 1c shows a scatter diagram of 100 points of (R^2, α). The points gather around the diagonal line, but tend to be above it: coefficient α tends to overestimate R^2.

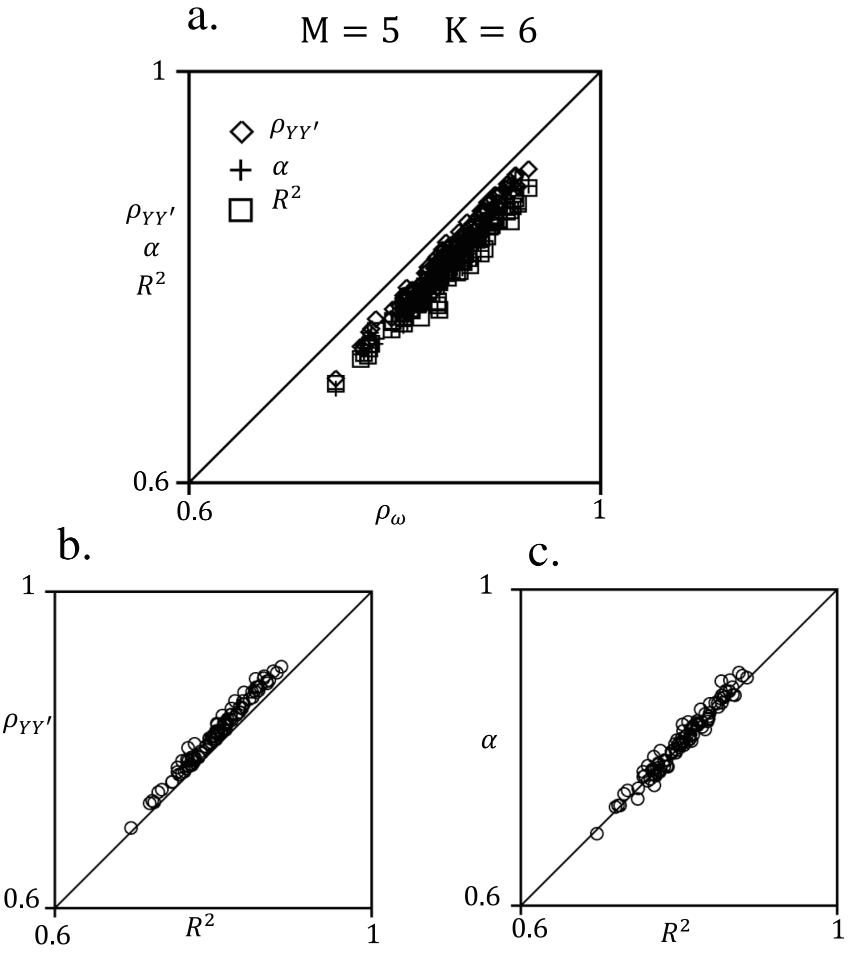

Figure 2. Scatter diagrams of points generated by 100 simulations for number of items $\mathrm{M} = 5$ and the number of categories $\mathrm{K} = 6$. In each simulation, parameters were sampled randomly with the following distributions: $\mu_j \sim U[-1, 1]$, $\lambda_j \sim U[0.5, 0.95]$, and $C_k \sim U[-1.6 + 0.5k, -1.4 + 0.5k]$, where $U[a, b]$ denotes a uniform distribution on an interval $[a, b]$. Figure 2a displays 100 sets of 3 points $(\rho_\omega, \rho_{YY'})$, (ρ_ω, α), and (ρ_ω, R^2). Figure 2b displays 100 points of $(R^2, \rho_{YY'})$. Figure 2c displays 100 points of (R^2, α).

When the number of categories increases from $K = 2$ to $K = 6$, coefficients become larger than that for $K = 2$ (Figures 1 and 2). Notice the differences in scales of the diagrams, in Figure 1 from 0.4 to 1 and in Figure 2 from 0.6. Points in Figure 2 have a closer approach to diagonal lines than those in

Figure 1, especially when considering the expansion of scales in Figure 2. However, general tendencies shown in Figure 2 are the same as in Figure 1. All of 3×100 points of $(\rho_\omega, \rho_{YY'})$, (ρ_ω, α), and (ρ_ω, R^2) are below the diagonal line (Figure 2a). Coefficient $\rho_{YY'}$ overestimates R^2 (Figure 2b), and α tends to overestimate R^2 (Figure 2c).

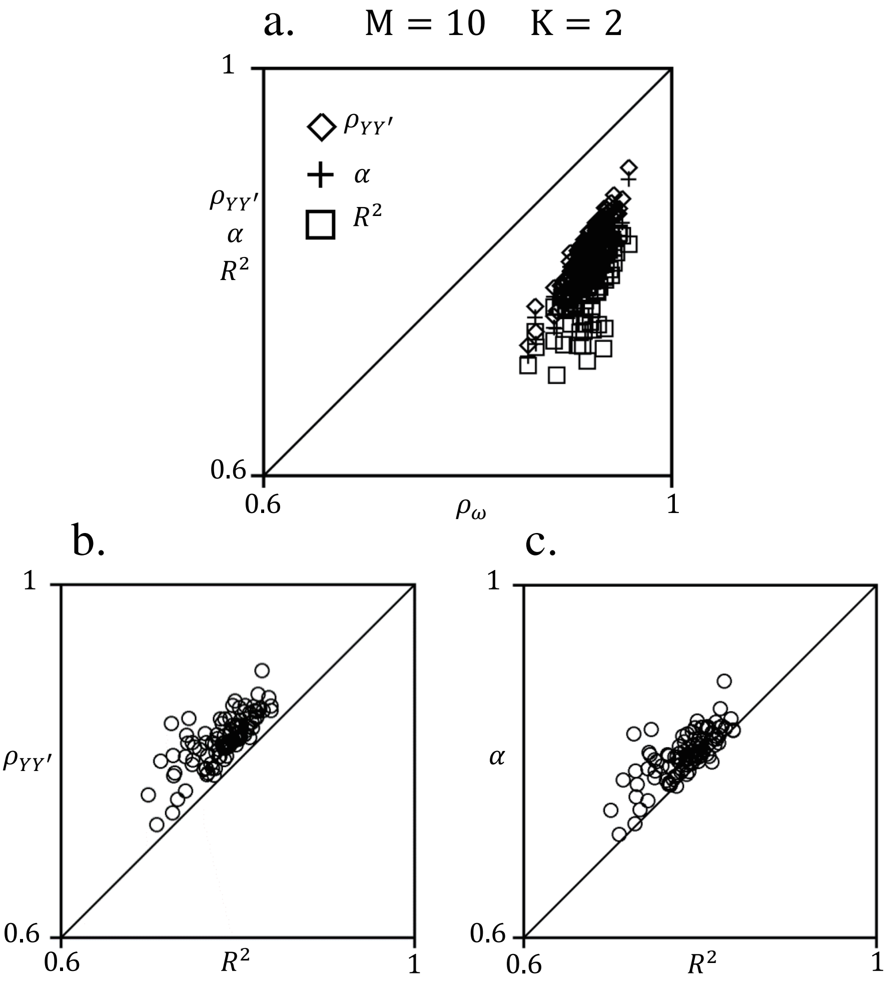

Figure 3. Scatter diagrams of points generated by 100 simulations for the number of items $\mathrm{M} = 10$ and the number of categories $\mathrm{K} = 2$. In each simulation, parameters were sampled randomly with the following distributions: $\mu_j \sim U[-1, 1]$, $\lambda_j \sim U[0.5, 0.95]$, and $C_1 \sim U[-0.5, 0.5]$, where $U[a, b]$ denotes a uniform distribution on an interval $[a, b]$. Figure 3a displays 100 sets of 3 points $(\rho_\omega, \rho_{YY'})$, (ρ_ω, α), and (ρ_ω, R^2). Figure 3b displays 100 points of $(R^2, \rho_{YY'})$. Figure 3c displays 100 points of (R^2, α).

When the number of items M increases from $M = 5$ to $M = 10$ with the number of categories $K = 2$, values of reliability coefficients increase (Figure 3). Notice that scales in Figure 3 are expanded; ranges in Figure 3 are from 0.6 to 1, while those in Figure 1 are from 0.4 to 1. Moreover, in Figure 3a, $\rho_{YY'}$, α, and R^2 are lower than the corresponding ρ_ω. Coefficient $\rho_{YY'}$ overestimates R^2

(Figure 3b). Coefficient α tends to overestimate R^2 more often for $M = 10$ than for $M = 5$ (compare Figure 3c and Figure 1c). Increasing the number of items stabilizes the tendency of overestimation of R^2 by α.

Increasing the number of items from $M = 5$ to $M = 10$ for the number of categories $K = 6$ increases values of reliability coefficients (Figure 4). Notice that scales in Figure 4 are from 0.8 to 1, while those in Figure 2 are from 0.6 to 1. Figure 4 shows that $\rho_{YY'}$, α, and R^2 are below corresponding ρ_ω (Figure 4a), $\rho_{YY'}$ overestimates R^2 (Figure 4b), and α tends to overestimate R^2 (Figure 4c). Comparison of Figure 4c and Figure 2c shows that for $K = 6$, increasing M from 5 to 10 intensifies the tendency of overestimation of R^2 by α.

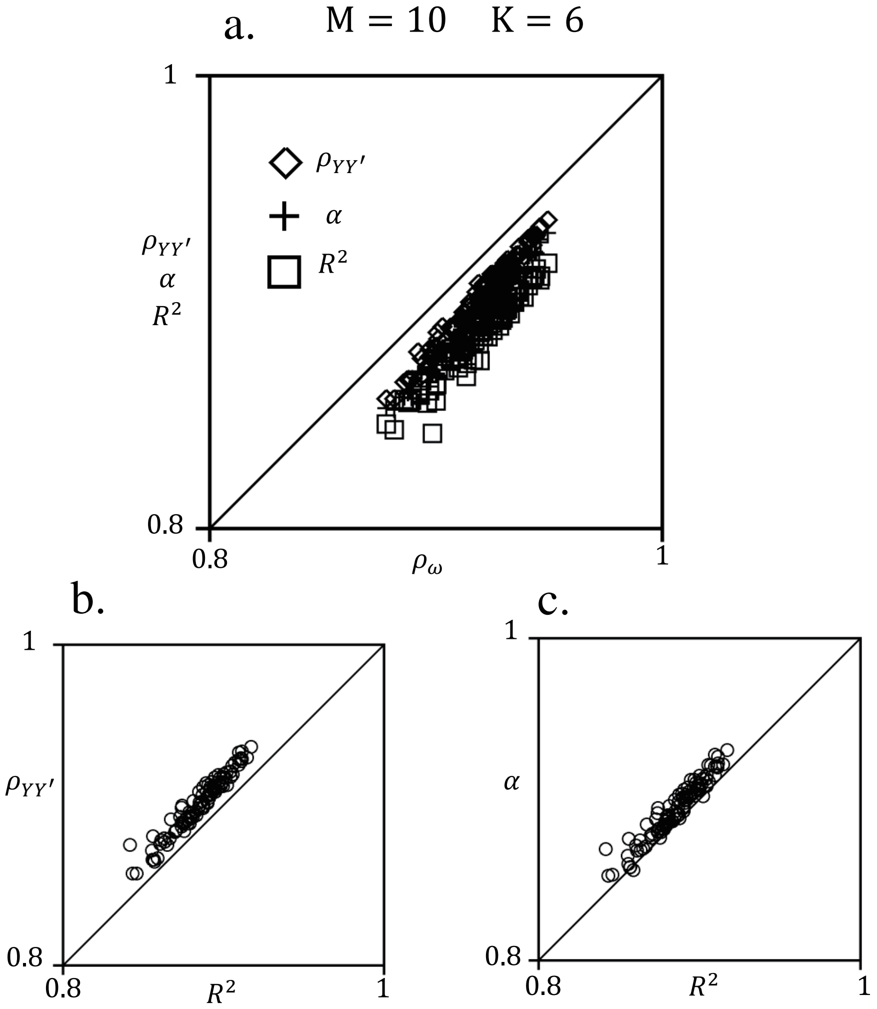

Figure 4. Scatter diagrams of points generated by 100 simulations for the number of items $\mathrm{M} = 10$ and the number of categories $\mathrm{K} = 6$. In each simulation, parameters were sampled randomly with the following distributions: $\mu_j \sim U[-1, 1]$, $\lambda_j \sim U[0.5, 0.95]$, and $C_k \sim U[-1.6 + 0.5k, -1.4 + 0.5k]$, where $U[a, b]$ denotes a uniform distribution on an interval $[a, b]$. Figure 4a displays 100 sets of 3 points $(\rho_\omega, \rho_{YY'})$, (ρ_ω, α), and (ρ_ω, R^2). Figure 4b displays 100 points of $(R^2, \rho_{YY'})$. Figure 4c displays 100 points of (R^2, α).

Discussion

The purpose of measurement is to estimate a true value of an attribute from an observed score. This framework of measurement for ordinal categorical items can be represented by Regression Model 11. The strength of the relation of variables Y and F in Model 11 can be represented by the coefficient of determination R^2 of Model 11. Simulation results show that $\rho_{YY'}$ overestimates R^2 (Figures 1b, 2b, 3b, and 4b). That is, actual relation R^2 of observed scores and true values are weaker than correlation coefficients $\rho_{YY'}$ of parallel tests. $\rho_{YY'}$ is also considered the reliability coefficient calculated by the test-retest method. Coefficient α tends to overestimate actual relation R^2 of observed scores and true values, which implies overestimation of R^2 by glb. This tendency becomes more stable as the number of items increases: compare Figures 3c and 4c with Figures 1c and 2c, respectively. It seems that increasing the number of items stabilizes the tendency of overestimation by coefficient α.

Today, many studies on test theory are conducted employing IRT and the theory that treats the sum of item scores is called classic test theory (CTT). However, in the case of ordinal categorical items, basic models in CTT are fundamentally the same as those in IRT, i.e., Thurstone’s model, which is employed in this study. Item characteristic curves (ICCs) of normal ogive models in IRT can be derived from Thurstone’s model. Basic models in IRT and CTT can be considered the same type, but differences between them can be observed. Basic models (e.g., Model 1) in CTT contain error terms as components and randomness of responses are explained by error terms. On the other hand, starting models in IRT are ICCs and do not contain error terms, and the randomness of responses is explained by the probabilistic nature of models, i.e., ICCs, which do not contain error terms. Hence, the concept of error in estimation in the case of IRT is not derived from the basic model, i.e., ICC, but is discussed in relation to the method of estimation of the attribute. For example, errors in maximum likelihood estimation (MLE) are evaluated (Samejima, 1994; Toyoda, 1989).

Another difference is that in CTT, measurement is given by the sum of observed scores, but in IRT, a value of the attribute is inferred from the observed pattern of item responses. That is, in CTT, measurement is given as an observed value, but in IRT, measurement is an inferred value from an observed pattern of responses based on the model.

Choice between the use of CTT and IRT would be made considering various conditions under which research studies are conducted. In 22 articles published by *The Japanese Journal of Psychology* in 2014, ordinal categorical items are used; 17 report reliability coefficients, all of which are coefficient alphas. These facts show that scales composed of ordinal categorical items are popular in psychology, at least in Japan, and CTT is chosen as a framework to analyze scales. In using CTT, coefficient alpha is very popular. However, results from our simulation show that coefficient alpha overestimates the actual relationship between a true value and an observed sum of scores on ordinal categorical items. The coefficient of determination R^2 should be used to indicate the actual relationship.

References

Cronbach, L. J. (1961). *Essentials of psychological testing, second edition.* New York: Harper Row and John Weatherhill, Inc.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont: Wadsworth Group/Thomson Learning.

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*, 155−167.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255−282.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151−160.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah: Lawrence Erlbaum Associates, Publishers.

Okamoto, Y. (2013). A direct Bayesian estimation of reliability. *Behaviormetrika*, *40*, 149−168.

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

Toyoda, H. (1989). The methods for estimating the reliability coefficient under item response model. *Japanese Journal of Educational Psychology, 37*, 283−285. (in Japanese with English abstract)