Japan Women's University Journal:
Faculty of Integrated Arts and Social Sciences, 2017, 28, 59-78.

Bayesian Estimation of Ordinal Categorical Items' Reliability Coefficients: Relationship Between True Values and Observed Scores

Yasuharu Okamoto

Abstract

A model and its associated algorithms are presented for an ordinal categorical items' reliability coefficient that denotes the predictive power of a regression of observed scores on true values. For categorical items, reliability coefficients given with respect to the three types of reliability, precision, consistency, and predictive power, mutually disagree in general. The predictive power reliability coefficient denotes the relationship between observed scores and true values, and other reliability coefficients overestimate the predictive power. The algorithms proposed for predictive reliability were applied to hypothetical datasets, and the results show that they are successful.

Keywords: reliability, predictive power, regression, ordered categorical items, Bayesian.

Bayesian Estimation of Ordinal Categorical Items' Reliability Coefficients: Relationship Between True Values and Observed Scores

順序カテゴリ項目尺度の信頼性係数のベイズ推定 —真値と観測値の関係—

OKAMOTO Yasuharu
岡 本 安 晴

[Abstract] A model and its associated algorithms are presented for an ordinal categorical items' reliability coefficient that denotes the predictive power of a regression of observed scores on true values. For categorical items, reliability coefficients given with respect to the three types of reliability, precision, consistency, and predictive power, mutually disagree in general. The predictive power reliability coefficient denotes the relationship between observed scores and true values, and other reliability coefficients overestimate the predictive power. The algorithms proposed for predictive reliability were applied to hypothetical datasets, and the results show that they are successful.

[Key Words] reliability, predictive power, regression, ordered categorical items, Bayesian.

1. Introduction

1.1 Three Aspects of Reliability

This study proposes a model and its associated estimation methods of ordinal categorical items' reliability coefficient that are more appropriate than current ones, e.g. alpha coefficient. Three aspects of reliability have been indicated–precision, consistency, and predictive power–and corresponding to each, a mathematical definition is formulated (cf. Raykov & Marcoulides, 2011). The proposed method estimates a reliability coefficient from the perspective of predictive power. But before discussing ordinal categorical items, we consider continuous items, which are usually assumed.

When reliability is considered to denote precision, an observed score is decomposed to a true score and an error, and reliability coefficient ρ is given by the following equation (McDonald, 1999; Crocker & Algina, 1986; Guilford, 1954):

$$\rho = \frac{Var(true\ score)}{Var(observed\ score)}, \tag{1}$$

where $Var(\cdot)$ denotes a variance. When an observed score X_{ij} on item j of person i is given by the following one-factor model:

$$X_{ij} = a_j + \lambda_j F_i + e_{ij} \tag{2}$$

ρ can be given by the following equation, denoted by ω (McDonald, 1999),

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

where M is the number of items, a_j a constant, λ_j a factor loading, F a common factor with variance 1 of the measured attribute, and e_j an error with variance ψ_j^2 independent of other variables.

From Equation 1, the popular coefficient alpha α can be derived (McDonald, 1999; Crocker & Algina, 1986):

$$\alpha = \frac{M}{M-1}\left\{1 - \frac{\sum_{j=1}^{M} Var(X_{ij})}{Var(X_i)}\right\}, \quad X_i = \sum_{j=1}^{M} X_{ij}.$$

As we know, α is equal to ω when all λ_js have the same value. Some explain that α is a lower bound of the reliability coefficient's estimates (Crocker & Algina, 1986; Fife, Mendoza & Terry, 2012; Furr & Bacharach, 2008; McDonald, 1999). The greatest lower bound (glb) is proposed as the reliability coefficient more adequate than α as a lower-bound estimate (Jackson & Agunwamba, 1977; Sijtsma, 2009; Ten Berge & Sočan, 2004). It is shown that $\alpha \leq glb$. However, α might be greater than the reliability coefficient's true value (Yang & Green, 2011), and simulations that obtained αs greater than true reliability coefficients are reported (Okamoto, 2013; Yang & Green, 2010).

When reliability is considered to denote consistency, the correlation coefficient between scores on parallel test forms is used (Crocker & Algina, 1986; Cronbach, 1961; Guttman, 1945). Two test forms X and X' are parallel when they are represented as follows:

$$X = T + E, \quad \text{and} \quad X' = T + E',$$

where T is a shared true score, and E and E' are errors with the same variances and independent of each other and T. In this case, reliability coefficient $\rho_{XX'}$ is given by

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

where $Cor(X, X')$ denotes the correlation coefficient of X and X'. As we know, $\rho_{XX'}$ is equal to ρ.

Reliability is also considered to denote predictive power, and a regression model is used. Lord and Novick (1968) regress an observed score on a true score. Raykov and Marcoulides (2011) emphasize

reliability's predictive power, saying "Reliability bears a distinctive relationship to the predictive power with which one can predict observed score from true score (p. 139)." They discuss two types of regression models: regression of an observed score on a true score and regression of a true score on an observed score. King, Rosopa, and Minium (2011) explain interchangeability between dependent and predictor variables. In both cases, strength of relationship, i.e., predictive power, is denoted by the coefficient of determination R^2-index, which is given by

$$R^2 = [Cor(observed\ score, true\ score)]^2.$$

As is known,

$$R^2 = \rho .$$

1.2. Ordinal Categorical Items

Traditionally, reliability coefficients are discussed for continuous items, and reliability coefficients defined for continuous items have been applied to ordinal categorical items. However, models for ordinal categorical items have been also proposed. Reliability for ordinal categorical items can also be discussed from the three perspectives. To clarify the discussion, a basic model of ordinal categorical items is needed. It is known by various names, e.g., a nonlinear structural equation model (SEM) (Green & Yang, 2009), an ordinal probit regression (Okamoto, 2013), or a Thurstonian item-response theory (IRT) model (Brown & Maydeu-Olivares, 2013). Raykov, Dimitrov, and Asparoukov (2010) indicate the equivalence of these models and the two-parameter normal ogive model in IRT. The models set in this study are as follows.

Let U_{ij} be an unobserved continuous value on item j of person i. A single factor model for U_{ij}, the same as that for X_{ij} (Equation 2), is set as follows:

$$U_{ij} = \mu_j + \lambda_j F_i + E_{ij}, \tag{3}$$

where F_i is a common factor with the standard normal distribution of person i, λ_j is a factor loading, μ_j is a position parameter, and E_{ij} is an error that has normal distribution with mean 0 and variance ψ_j^2. Error terms E_{ij}s are independent of each other and F_i. The observed response Y_{ij} on item j, based on U_{ij}, is made according to the rule

$$Y_{ij} = k, \quad \text{if } C_{k-1} \leq U_{ij} < C_k, \tag{4}$$

where C_ks are category boundaries for response Y_{ij}, and

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

An observed score Y_i can be obtained as the sum of Y_{ij}s, that is,

$$Y_i = \sum_{j=1}^{M} Y_{ij} \,. \tag{5}$$

In the case of ordinal categorical items, an observed continuous variable X_{ij} of Model 2 becomes unobserved and is denoted by U_{ij}. Unobserved variable U_{ij} is categorized by Y_{ij}, which is observed.

Under the model shown above for ordinal categorical items, reliability coefficients can be discussed as follows.

1.2.1. Current reliability coefficients for categorical items

Reliability coefficient ρ_ω, which denotes a test's precision as coefficient ω for underlying latent variables U_{ij}s, is proposed by Okamoto (2013). That is, ρ_ω is given by

$$\rho_\omega = \frac{\left(\sum_{j=1}^{M} \lambda_j\right)^2}{\left(\sum_{j=1}^{M} \lambda_j\right)^2 + \sum_{j=1}^{M} \psi_j^2}. \tag{6}$$

Parameters λ_js and ψ_js are estimated based on samples from Markov chain Monte Carlo (MCMC), which is a popular method in Bayesian analysis. ρ_ω denotes latent variables' precision and is independent of categories used, so it can be considered an intrinsic character of items. But, ρ_ω does not denote precision of the observed sum of ordinal categorical items.

Reliability coefficient, explained to denote precision of ordinal categorical items, is proposed for binary items (Dimitrov, 2003; Raykov, Dimitrov, & Asparouhov, 2010). A binary score Y_{ij} is decomposed as a sum of a true value τ_{ij} and an error e_{ij}, i.e.,

$$Y_{ij} = \tau_{ij} + e_{ij} \,. \tag{7}$$

When we interpret Equation 7 as a sum of a true value and an error, the meaning of "true value" is ambiguous. A score on a binary item takes one of two values, 1 or 2, for example. Which value is the true one? Dimitrov (2003) derives the true value as $1 + P_j(\theta_i)$, where $P_j(\theta_i)$ denotes probability of response $Y_{ij} = 2$ on item j of person i with a trait of score θ_i. (He sets a score on a binary item as 0 or 1.) Hence,

$$1 < \tau_{ij} = P_j(\theta_i) < 2 \,.$$

However, a score on binary items is 1 or 2, so true value τ_{ij} should also be 1 or 2, i.e., one of the binary responses. Hence, the concept of true value τ_{ij} in Model 7 is contradictory, or at least ambiguous, although taking formal operation of expectation of τ_{ij} obscures the contradiction. When we measure a concept, we want to know its strength, the true value of which is reflected by F_i of Equation 3, so precision of observed scores should be considered with respect to F_i of Equation 3.

Coefficient alpha α is also calculated for ordinal categorical items as follows:

$$\alpha = \frac{M}{M-1}\left\{1 - \frac{\sum_{j=1}^{M} Var(Y_{ij})}{Var(Y_i)}\right\},$$

which is obtained by replacing X_{ij}s for continuous items with Y_{ij}s for ordinal categorical items.

When considering a test's consistency, the correlation coefficient of parallel test forms is used as a reliability coefficient (Green & Young, 2009). A parallel test Y' of Y is represented as follows:

$$U'_{ij} = \mu_j + \lambda_j F_i + E'_{ij} \tag{8}$$

$$Y'_{ij} = k, \quad \text{if } C_{k-1} \leq U'_{ij} < C_k, \tag{9}$$

$$Y'_i = \sum_{j=1}^{M} Y'_{ij}, \tag{10}$$

where Equations 8, 9, and 10 correspond to Equations 3, 4, and 5, respectively. The same symbols μ_j, λ_j, F_i, and C_k denote shared values, respectively. Parallel tests differ only in error terms independent of other variables and have the same variance as corresponding error terms. The model for parallel tests can be applied to the test-retest method. In this study, the reliability coefficient defined by the correlation coefficient of parallel tests Y and Y' is denoted by $\rho_{YY'}$:

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

1.2.2. Proposed predictive reliability coefficient

Predictive power for continuous items is denoted by the R^2-index of a regression equation of observed scores on true scores, and is equal to ρ. For ordinal categorical items, an observed score is Y_i, and a true score is F_i, which reflects the strength of an attribute of a concept to be measured. Hence, the regression equation is represented as follows:

$$Y_i = a + bF_i + e \tag{11}$$

Values of a and b are determined by the least-squares criterion. As is known, R^2-index is given by the squared correlation of Y_i and F_i, so the reliability coefficient as predictive power, denoted by ρ_{cat} in this study, is given by

$$\rho_{cat} = [Cor(Y_i, F_i)]^2.$$

Although Raykov and Marcoulides (2011) emphasized distinctive relationship of reliability and the predictive power, they did not give an explicit definition of reliability for categorical items from the point of view of predictive power. As an approximate method, they proposed parceling of categorical items. In contrast to their approximate method, ρ_{cat} proposed above is defined explicitly as the predictive power of Regression 11. Equations to calculate ρ_{cat} are given in the Appendix.

Reliability coefficient ρ_ω also denotes the predictive power of a regression of U_i on F_i. i.e., ρ_ω is equal to R^2-index of regression equation

$$U_i = a_U + b_U F_i + e_U, \quad U_i = \sum_{j=1}^{M} U_{ij},$$

where regression coefficients a_U and b_U are determined by the least-squares criterion.

Values of ρ_ω, $\rho_{YY'}$, α, and ρ_{cat} are determined from parameters μ_js. λ_js. and ψ_js. Reliability coefficient ρ_ω is given by Equation 6, and $\rho_{YY'}$, α, and ρ_{cat} are given by equations listed in the Appendix.

Comparison of reliability coefficients for ordinal categorical items was done for concrete values of parameters and is reported in the next section.

1.3. Comparison of Reliability Coefficients for Ordinal Categorical Items

Given the values of parameters λ_js and ψ_js, reliability coefficient ρ_ω can be calculated by Equation 6. Other coefficients–ρ_{cat}, $\rho_{YY'}$, and α–can be calculated using equations listed in the Appendix from values of μ_js, λ_js, ψ_js, and C_ks. For the following combination of parameter values, comparisons of reliability coefficients ρ_{cat}, ρ_ω, $\rho_{YY'}$, and α were conducted. Effects of the number of categories K and number of items M on reliability coefficients were investigated with parameter values $\lambda_j = 0.7$, $\psi_j^2 = 1 - \lambda_j^2$ (i.e., parallel items [McDonald, 1999]), and $C_1 = 0$ for $K = 2$; $C_1 = -0.2$, $C_2 = 0.2$ for $K = 3$; $C_k = -1.5 + 0.5k$, $k = 1, \cdots, 5$ for $K = 6$; $C_k = -1.875 + 0.375\mathrm{k}$, $k = 1, \cdots, 9$ for $K = 10$. Parameters μ_js are set at 0 for all combinations. Figure 1 shows the values of reliability coefficients ρ_ω, $\rho_{YY'}$, α, and ρ_{cat} for number of items $M = 5$ (a) and for $M = 10$ (b). Reliability coefficient ρ_ω is constant over the number of categories K, because ρ_ω is a reliability coefficient for latent variables U_{ij}s and is not affected by category boundaries C_ks (Equation 6). On the other hand, ρ_{cat} varies with K. Reliability coefficient ρ_{cat} is substantially lower than ρ_ω for small K; this means that the amount of lost information about F_i contained in U_{ij} increases as the number of categories decreases. The difference $\rho_\omega - \rho_{cat}$ can be interpreted as the amount of information lost by discretization of the continuous variable U_{ij} to the ordinal categorical variable Y_{ij} (Equation 4). Figure 1 shows that for the number of categories $K = 10$, this loss of information becomes negligible, that is, Y_{ij}s with $K = 10$ can be considered approximately continuous items. Over the range of K, from 2 to 10, reliability coefficients $\rho_{YY'}$ and α are between ρ_ω and ρ_{cat}. They underestimate reliability ρ_ω for latent variables U_{ij}s, and overestimate reliability coefficient ρ_{cat}, which represents the actual relation's strength between a true value F_i and an observed value Y_i. Other simulations by the author showed a similar tendency. Reliability coefficient ρ_{cat} denotes the strength of relation between true values and observed scores, and reliability coefficients other than ρ_{cat} tend to overestimate ρ_{cat}. Hence, estimating ρ_{cat} to know the strength of relation between true values and observed scores is worthwhile. The next section proposes Bayesian methods of estimation for reliability coefficient ρ_{cat}.

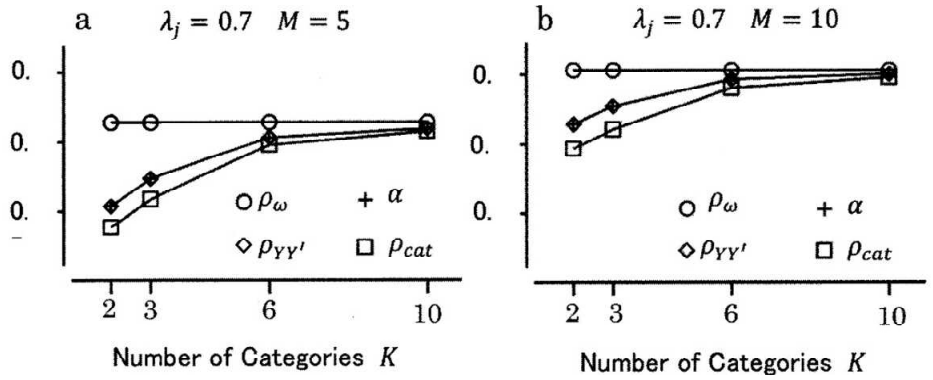

Figure 1. Reliability coefficients ρ_ω, $\rho_{YY'}$, α, and ρ_{cat}. Number of items M is 5 (a) or 10 (b). Factor loadings $\lambda_j s$ are 0.7, and variances of errors $\psi_j^2 s$ are $1 - \lambda_j^2$. Category boundary is $C_1 = 0$ for number of categories $K = 2$, and category boundaries are $C_1 = -0.2$, and $C_2 = 0.2$ for $K = 3$; $C_k = -1.5 + 0.5k, k = 1, \cdots, 5$ for $K = 6$; or $C_k = -1.875 + 0.375k, k = 1, \cdots, 9$ for $K = 10$. All μ_js are 0.

2. Bayesian Algorithms

Because of differences in constraints set to identify the values of parameters between binary items and items with more than two categories, algorithms are presented separately for each type of item.

2.1. Items with More than Two Categories

Set

$$\boldsymbol{F} = (F_1, , F_N), \boldsymbol{\mu} = (\mu_1, \cdots, \mu_M), \boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_M), \boldsymbol{\psi} = (\psi_1, \cdots, \psi_M), \boldsymbol{C} = (C_0, \cdots, C_K),$$

where N is the number of persons. Corresponding to these parameters, the probability of response $Y_{ij} = k$ of person i on item j is given by

$$P(Y_{ij} = k \,|\boldsymbol{F}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{C}) = P(C_{k-1} \le U_{ij} < C_k \,|\boldsymbol{F}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{C})$$

$$= \Phi\left(\frac{C_k - (\mu_j + \lambda_j F_i)}{\psi_j}\right) - \Phi\left(\frac{C_{k-1} - (\mu_j + \lambda_j F_i)}{\psi_j}\right) \tag{12}$$

Set

$$\boldsymbol{Y} = (Y_{11}, Y_{12}, \cdots, Y_{NM}),$$

where N is the number of persons.

Under the independence assumption, we have the following likelihood function:

$$L(\boldsymbol{F},\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\psi},\boldsymbol{C}\,|\boldsymbol{Y}) = \prod_{i=1}^{N}\prod_{j=1}^{M} P(Y_{ij}\,|\boldsymbol{F},\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\psi},\boldsymbol{C}).$$

Hence, the posterior probability function is given by

$$P(\boldsymbol{F},\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\psi},\boldsymbol{C}\,|\boldsymbol{Y}) \propto P_0(\boldsymbol{F},\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\psi},\boldsymbol{C})\prod_{i=1}^{N}\prod_{j=1}^{M} P(Y_{ij}\,|\boldsymbol{F},\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\psi},\boldsymbol{C}), \tag{13}$$

where $P_0(\boldsymbol{F},\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\psi},\boldsymbol{C})$ is a prior probability function and is given in this study by

$$P_0(\boldsymbol{F},\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\psi},\boldsymbol{C}) = P_0(\boldsymbol{F})P_0(\boldsymbol{\mu})P_0(\boldsymbol{\lambda})P_0(\boldsymbol{\psi})P_0(\boldsymbol{C})$$

$$P_0(\boldsymbol{F}) = \prod_{i=1}^{N} P(F_i), P_0(\boldsymbol{\mu}) = \prod_{j=1}^{M} P(\mu_j), P_0(\boldsymbol{\lambda}) = \prod_{j=1}^{M} P(\mu_j), P_0(\boldsymbol{\psi}) = \prod_{j=1}^{M} P(\psi_j).$$

The prior distribution for F_i is the standard normal distribution, and prior distributions of μ_j, λ_j and ψ_j are chosen similarly to those of Okamoto (2013), i.e., uniform distributions on sufficiently large regions with restrictions $\lambda_j > 0$ and $\psi_j > 0$. Prior distribution $P_0(\boldsymbol{C})$ for C_ks is set to be uniformly distributed on sufficiently large regions under the constraint

$$C_1 < \cdots < C_{K-1}.$$

That is,

$$\mathrm{P}(C_1,\cdots,C_{K-1}) = \begin{cases} constant & if\ C_1 < \cdots < C_{K-1} \\ 0 & otherwise \end{cases}$$

Equation 12 implies that an origin and a unit of the scale are arbitrary. For the origin and the unit, Okamoto (2013) sets

$$C_1 = -1 \text{ and } C_{K-1} = 1, \tag{14}$$

to estimate ρ_ω, because Equation 6 is invariant when units are rescaled by the same factor among items or origins are shifted, not necessarily by the same distance for each item. To identify parameter values, the origin and unit are chosen by setting Constraint 14. Constraint 14 assumes that category boundaries are common among items and under this assumption reliability coefficient ρ_ω is given by Equation 6.

Under Constraint 14, μ_j can be interpreted to correspond to the modal category of item j. As μ_j becomes larger from a value smaller than $C_1 = -1$ to a value larger than $C_{K-1} = 1$, the distribution of categorical responses Y_{ij}s shifts from the left to the right.

Point estimates and posterior distributions of reliability coefficients can be estimated by the following Algorithm 1.

Algorithm 1 (Items with more than two categories: The program is available from the author.)

Step 1. Generate N_{main} samples from the posterior distribution (Equation 13) by Markov chain Monte

Carlo (MCMC). Denote the sample at time t by $\boldsymbol{\mu}^{(t)}$, $\boldsymbol{\lambda}^{(t)}$, $\boldsymbol{\psi}^{(t)}$, and $\boldsymbol{C}^{(t)}$. The value of N_{main} should be sufficiently large so that the point estimates of $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$, $\boldsymbol{\psi}$, and $\boldsymbol{C}$ are stable. In the program by the author of this study, N_{main} is set at 12,000.

Step 2. Calculate means of N_{main} samples generated in Step 1 as point estimates of parameters $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$, $\boldsymbol{\psi}$, and $\boldsymbol{C}$. Estimated values are denoted by parameters with carets above, e.g., $\hat{\mu}_j$, $\hat{\lambda}_j$, $\hat{\psi}_j$, and $\hat{C}_k$.

Step 3. Calculate ρ_{cat} using the point estimates of parameters. Other reliability coefficients, e.g., ρ_{ω}, $\rho_{YY'}$, or α, can also be calculated from the parameters' point estimates. Equations to calculate reliability coefficients from parameters are listed in the Appendix.

If you want to obtain reliability coefficients' posterior distributions, go to the next step.

Step 4. As the number of samples is too large, calculation of the posterior distribution of reliability coefficient ρ_{cat} requires too much time. To reduce computation time, randomly select N_{sub} ($< N_{main}$) samples from the samples generated in Step 1. The author's program sets N_{sub} =1,000. Denote the sth sample of N_{sub} samples from N_{main} samples by $\boldsymbol{\mu}^{(s/sub)}$, $\boldsymbol{\lambda}^{(s/sub)}$, $\boldsymbol{\psi}^{(s/sub)}$, and $\boldsymbol{C}^{(s/sub)}$.

Step 5. Calculate the sth sample of reliability coefficient $\rho_{cat}^{(s)}$ as that determined by the sth sample of parameters $\boldsymbol{\mu}^{(s/sub)}$, $\boldsymbol{\lambda}^{(s/sub)}$, $\boldsymbol{\psi}^{(s/sub)}$, and $\boldsymbol{C}^{(s/sub)}$. If you are also interested in other reliability coefficients, e.g., $\rho_{YY'}^{(s)}$, $\alpha^{(s)}$, or $\rho_{\omega}^{(s)}$, they can also be calculated.

Step 6. Estimate the reliability coefficient's posterior distribution by samples $\rho_{cat}^{(s)}$s. Posterior distributions of other reliability coefficients can also be estimated from samples $\rho_{YY'}^{(s)}$s, $\alpha^{(s)}$s, or $\rho_{\omega}^{(s)}$s.

The sampling method in Step 1 and used in the author's program is called Metropolis-within-Gibbs (Robert & Casella, 2010a, 2010b) or the component-wise version of the Metropolis-Hastings algorithm (Gamerman & Lopes, 2006) and is essentially the same as that by Okamoto (2013). Hence, details of MCMC algorithms used in this study are omitted, but see Okamoto (2013) for details.

2.2 Binary Items

For binary items, Constraint 14 cannot be employed, because a single category boundary C_1 bisects the continuum of the latent variable U_{ij}. Okamoto (2013) uses the constraint $C_1 = 0$ under the restriction of essentially tau-equivalent items, i.e.,

$$\lambda_1 = \cdots = \lambda_m = 1. \tag{15}$$

But, Restriction 15 is not needed to estimate ρ_{cat}, $\rho_{YY'}$, or α, because these values can be calculated from probability $P(Y_{ij} = k|\boldsymbol{F}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{C})$ (Equation 12) using equations listed in the Appendix. In the case

of binary items, Equation 12 becomes

$$P(Y_{ij}=1|\boldsymbol{F},\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\psi},\boldsymbol{C})=\Phi\left(\frac{C_1-(\mu_j+\lambda_jF_i)}{\psi_j}\right).$$

Because of arbitrariness of origin, set

$$C_1=0,$$

then we have

$$P(Y_{ij}=1|\boldsymbol{F},\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\psi},\boldsymbol{C})=\Phi\left(\frac{-(\mu_j+\lambda_jF_i)}{\psi_j}\right)=\Phi\left(-\left\{\frac{\mu_j}{\psi_j}+\frac{\lambda_j}{\psi_j}F_i\right\}\right).$$

Put

$$\mu_{bj}=\frac{\mu_j}{\psi_j}\ \text{and}\ \lambda_{bj}=\frac{\lambda_j}{\psi_j},$$

and then probabilities for binary items can be given by parameters μ_{bj}s and λ_{bj}s, instead of μ_js, λ_js, and ψ_js. That is, for binary items, the unit for each item j is set so that $\psi_{bj}^2=Var(E_j)=1$, and the origin is set so that $C_1=0$. Hence, we can estimate reliability coefficients ρ_{cat}, $\rho_{YY'}$, or α for binary items, which do not need to satisfy the condition of essentially tau-equivalent items. But, reliability coefficient ρ_ω cannot be calculated from these parameters because item js do not have the same common unit.

For binary items, the stochastic model can be written as follows:

Set

$$\boldsymbol{F}=(F_1,,F_N),\boldsymbol{\mu}_b=(\mu_{b1},\cdots,\mu_{bM}),\boldsymbol{\lambda}_b=(\lambda_{b1},\cdots,\lambda_{bM}),\boldsymbol{Y}=(Y_{11},Y_{12},\cdots,Y_{NM}).$$

Probabilities of Y_{ij} are given as follows,

$$P(Y_{ij}=1\mid\boldsymbol{F},\boldsymbol{\mu}_b,\boldsymbol{\lambda}_b)=\Phi\left(-(\mu_{bj}+\lambda_{bj}F_i)\right)$$

and

$$P(Y_{ij}=2\mid\boldsymbol{F},\boldsymbol{\mu}_b,\boldsymbol{\lambda}_b)=1-P(Y_{ij}=1\mid\boldsymbol{F},\boldsymbol{\mu}_b,\boldsymbol{\lambda}_b).$$

Then, we have the following likelihood function:

$$L(\boldsymbol{F},\boldsymbol{\mu}_b,\boldsymbol{\lambda}_b\,|\boldsymbol{Y})=\prod_{i=1}^{N}\prod_{j=1}^{M}P(Y_{ij}\mid\boldsymbol{F},\boldsymbol{\mu}_b,\boldsymbol{\lambda}_b).$$

The posterior distribution function is given by

$$P(\boldsymbol{F}, \boldsymbol{\mu}_b, \boldsymbol{\lambda}_b \,|\boldsymbol{Y}) \propto P_{b0}(\boldsymbol{F}, \boldsymbol{\mu}_b, \boldsymbol{\lambda}_b) \prod_{i=1}^{N} \prod_{j=1}^{M} P\left(Y_{ij} \mid \boldsymbol{F}, \boldsymbol{\mu}_b, \boldsymbol{\lambda}_b\right). \tag{16}$$

The prior distribution function $P_{b0}(\boldsymbol{F}, \boldsymbol{\mu}_b, \boldsymbol{\lambda}_b)$ is given in this study by

$$P_{b0}(\boldsymbol{F}, \boldsymbol{\mu}_b, \boldsymbol{\lambda}_b) = P_{b0}(\boldsymbol{F}) P_{b0}(\boldsymbol{\mu}_b) P_{b0}(\boldsymbol{\lambda}_b),$$

$$P_{b0}(\boldsymbol{F}) = \prod_{i=1}^{N} P(F_i), P_{b0}(\boldsymbol{\mu}_b) = \prod_{j=1}^{M} P\left(\mu_{bj}\right), P_{b0}(\boldsymbol{\lambda}_b) = \prod_{j=1}^{M} P\left(\lambda_{bj}\right).$$

The prior distribution for F_i is the standard normal distribution, and prior distributions of μ_{bj} and λ_{bj} are uniform distributions on sufficiently large regions with restriction $\lambda_{bj} > 0$.

Point estimates and reliability coefficients' posterior distributions can be estimated by the following Algorithm 2.

Algorithm 2 (Binary items: The program is available from the author.).

Step 1. Generate N_{main} samples from the posterior distribution (Equation 16) by Markov chain Monte Carlo (MCMC). Denote the sample at time t by $\boldsymbol{\mu}_b^{(t)}$ and $\boldsymbol{\lambda}_b^{(t)}$. The value of N_{main} should be sufficiently large so that the point estimates of $\boldsymbol{\mu}_b$ and $\boldsymbol{\lambda}_b$ are stable. In the author's program, N_{main} is set at 12,000.

Step 2. Calculate means of N_{main} samples generated in Step 1 as point estimates of parameters $\boldsymbol{\mu}_b$ and $\boldsymbol{\lambda}_b$. Estimated values are denoted by parameters with carets above, e.g., $\hat{\mu}_{bj}$ and $\hat{\lambda}_{bj}$.

Step 3. Calculate ρ_{cat} using point estimates of parameters. Other reliability coefficients, e.g., ρ_{ω}, $\rho_{YY'}$, or α, can also be calculated from parameters' point estimates. Equations to calculate reliability coefficients from parameters are listed in the Appendix. In applying equations in the Appendix to binary items, μ_j, λ_j, ψ_j, and C_1 are replaced by μ_{bj}, λ_{bj}, 1, and 0, respectively.

If you want to obtain reliability coefficients' posterior distributions, go to the next step.

Step 4. The number of samples is too large, so calculation of the posterior distribution of reliability coefficient ρ_{cat} requires too much time. To reduce computation time, randomly select N_{sub} ($< N_{main}$) samples from the samples generated in Step 1. The author's program sets N_{sub} =1,000. Denote the sth sample of N_{sub} samples from N_{main} samples by $\boldsymbol{\mu}_b^{(s/sub)}$ and $\boldsymbol{\lambda}_b^{(s/sub)}$.

Step 5. Calculate the sth sample of reliability coefficient $\rho_{cat}^{(s)}$ as that determined by the sth sample of parameters $\boldsymbol{\mu}_b^{(s/sub)}$ and $\boldsymbol{\lambda}_b^{(s/sub)}$. If you are also interested in other reliability coefficients, e.g., $\rho_{YY'}^{(s)}$, $\alpha^{(s)}$, or $\rho_{\omega}^{(s)}$, they can be calculated.

Step 6. Estimate the reliability coefficient's posterior distribution by samples $\rho_{cat}^{(s)}$s. Posterior distributions of other reliability coefficients can also be estimated from samples $\rho_{YY'}^{(s)}$s and $\alpha^{(s)}$s.

The sampling method in Step 1 and used in the author's program is essentially the same as

that used by Okamoto (2013). Hence, details of MCMC algorithms used in this study are omitted, but see Okamoto (2013) for details.

Examples of applications of the above algorithms to hypothetical datasets are presented in the next section.

3. Applications

Algorithms 1 and 2 were applied to hypothetical datasets of items with six categories (Table 1) and binary items (Table 2), respectively.

TABLE 1

A hypothetical dataset of 500 people on 10 items with 6 categories. The dataset was generated using Equations 3 and 4 with parameter values $\mu_1 = \cdots = \mu_5 = 0.1$, $\mu_6 = \cdots = \mu_{10} = -0.1$, $\lambda_j = \lambda_{j+5} = 0.525 + 0.075j$, $\psi_j^2 = 1 - \lambda_j^2$, $C_1 = -1.5$, $C_2 = -0.5$, $C_3 = 0$, $C_4 = 0.5$, $C_5 = 1.5$. The complete dataset is available from the author.

People	Itm1	Itm2	Itm3	Itm4	Itm5	Itm6	Itm7	Itm8	Itm9	Itm10
1	2	4	4	2	2	5	4	1	3	3
2	2	5	3	2	3	3	1	2	4	3
3	6	4	3	2	2	5	4	3	2	3
4	2	2	2	2	2	1	1	2	2	2
5	5	5	5	4	6	5	6	5	6	6
•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•
496	2	2	2	2	2	1	2	1	1	1
497	2	5	2	3	2	4	1	3	2	3
498	2	2	1	2	2	1	2	1	2	2
499	4	3	3	2	2	2	4	2	1	1
500	4	2	2	2	3	2	1	2	2	2

TABLE 2

A hypothetical dataset of 500 people on 10 binary items. The dataset was generated using Equations 3 and 4 with parameter values $\mu_1 = \cdots = \mu_5 = 0.1$, $\mu_6 = \cdots = \mu_{10} = -0.1$, $\lambda_j = \lambda_{j+5} = 0.525 + 0.075j$, $\psi_j^2 = 1 - \lambda_j^2$, $C_1 = 0$. The complete dataset is available from the author

People	Itm1	Itm2	Itm3	Itm4	Itm5	Itm6	Itm7	Itm8	Itm9	Itm10
1	1	2	2	1	1	2	2	1	1	1
2	1	2	1	1	1	1	1	1	2	1
3	2	2	1	1	1	2	2	1	1	1
4	1	1	1	1	1	1	1	1	1	1
5	2	2	2	2	2	2	2	2	2	2
•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•
496	1	1	1	1	1	1	1	1	1	1
497	1	2	1	1	1	2	1	1	1	1
498	1	1	1	1	1	1	1	1	1	1
499	2	1	1	1	1	1	2	1	1	1
500	2	1	1	1	1	1	1	1	1	1

The Table 1 dataset contains responses by 500 persons to 10 items with 6 categories. Point estimates in Step 2 of parameters for the Table 1 dataset were as follows:

$\hat{\mu}_1 \approx 0.033$, $\hat{\mu}_2 \approx 0.060$, $\hat{\mu}_3 \approx 0.074$, $\hat{\mu}_4 \approx 0.054$, $\hat{\mu}_5 \approx 0.023$, $\hat{\mu}_6 \approx -0.107$, $\hat{\mu}_7 \approx -0.082$, $\hat{\mu}_8 \approx -.084$, $\hat{\mu}_9 \approx -0.078$, $\hat{\mu}_{10} \approx -0.108$, $\hat{\lambda}_1 \approx 0.377$, $\hat{\lambda}_2 \approx 0.471$, $\hat{\lambda}_3 \approx 0.498$, $\hat{\lambda}_4 \approx 0.530$, $\hat{\lambda}_5 \approx 0.590$, $\hat{\lambda}_6 \approx 0.410$, $\hat{\lambda}_7 \approx 0.414$, $\hat{\lambda}_8 \approx 0.485$, $\hat{\lambda}_9 \approx 0.549$, $\hat{\lambda}_{10} \approx 0.586$,
$\hat{\psi}_1 \approx 0.510$, $\hat{\psi}_2 \approx 0.465$, $\hat{\psi}_3 \approx 0.460$, $\hat{\psi}_4 \approx 0.391$, $\hat{\psi}_5 \approx 0.276$, $\hat{\psi}_6 \approx 0.526$, $\hat{\psi}_7 \approx 0.509$, $\hat{\psi}_8 \approx 0.458$, $\hat{\psi}_9 \approx 0.376$, $\hat{\psi}_{10} \approx 0.315$, $C_1 = -1$ (Constraint 14), $\hat{C}_2 \approx -0.338$, $\hat{C}_3 \approx -0.001$, $\hat{C}_4 \approx 0.340$, $C_5 = 1$ (Constraint 14).

Considering randomness in generating sample data, and differences in category boundaries, $C_1 = -1.5$ and $C_5 = 1.5$ in generation of the dataset and $C_1 = -1$ and $C_5 = 1$ in estimation with Constraint 14, these point estimates correspond well to parameter values used in generation of the dataset Table 1.

From these point estimates, reliability coefficients were estimated in Step 3 as follows:

$$\hat{\rho}_{cat} \approx 0.905,\ \hat{\alpha} \approx 0.913,\ \hat{\rho}_{YY'} \approx 0.915,\ \hat{\rho}_{\omega} \approx 0.927.$$

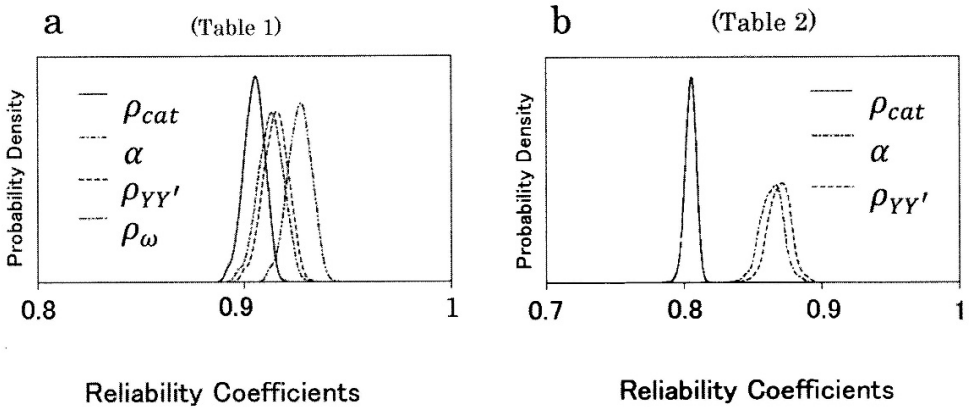

Figure 2. Posterior distributions of reliability coefficients for the data sets Table 1(a) and Table 2 (b). Each of the curves was drawn by kernel smoothing of 1000 sample points.

Posterior distributions of reliability coefficients estimated in Step 6 are shown in Figure 2a. Although the distributions rather overlapped, they shifted slightly from the left to the right in the order of ρ_{cat}, α, $\rho_{YY'}$, and ρ_{ω}.

For the Table 2 dataset (binary items), point estimates of parameters in Step 2 were as follows:

$\hat{\mu}_{b1} \approx 0.059$, $\hat{\mu}_{b2} \approx 0.147$, $\hat{\mu}_{b3} \approx 0.130$, $\hat{\mu}_{b4} \approx 0.125$, $\hat{\mu}_{b5} \approx 0.230$, $\hat{\mu}_{b6} \approx -0.224$, $\hat{\mu}_{b7} \approx -0.180$, $\hat{\mu}_{b8} \approx -0.172$, $\hat{\mu}_{b9} \approx -0.197$, $\hat{\mu}_{b10} \approx -0.449$, $\hat{\lambda}_{b1} \approx 0.806$, $\hat{\lambda}_{b2} \approx 1.112$, $\hat{\lambda}_{b3} \approx 1.092$, $\hat{\lambda}_{b4} \approx 1.363$, $\hat{\lambda}_{b5} \approx 3.273$, $\hat{\lambda}_{b6} \approx 0.886$, $\hat{\lambda}_{b7} \approx 0.721$, $\hat{\lambda}_{b8} \approx 1.018$, $\hat{\lambda}_{b9} \approx 1.589$, $\hat{\lambda}_{b10} \approx 2.212$.

Note that $\mu_{bj} = \mu_j/\psi_j$ and $\lambda_{bj} = \lambda_j/\psi_j$.

Considering randomness in generating sample data, and differences in units of scales, different ψ_js in generation of the dataset and the same ψ_js (the constraint $\psi_j = 1$ to set units) in estimation, these point estimates correspond well to parameter values used in generation of the dataset Table 2.

From these point estimates, reliability coefficients were calculated as follows:

$$\hat{\rho}_{cat} \approx 0.806,\ \hat{\alpha} \approx 0.866, \text{ and } \hat{\rho}_{YY'} \approx 0.871.$$

Since a common unit is not used for binary items, ρ_{ω} was not calculated.

Posterior distributions of the reliability coefficients estimated in Step 6 are shown in Figure 2b. The distribution of ρ_{cat} is clearly separated from those of α and $\rho_{YY'}$. Hence, estimating ρ_{cat} as the index that denotes binary items' predictive power is worthwhile. Reliability coefficients α and $\rho_{YY'}$ overestimate the relationship between true values and observed values.

4. Discussion

Three aspects of reliability—precision, consistency, and predictive power—were indicated, and reliability coefficients' corresponding definitions were discussed. These reliability coefficients are defined for continuous items and have the same value under their defined conditions. Hence, they can be considered to denote the same concept, that is, reliability from different perspectives.

In psychology, however, most scales are constructed essentially by the sum of scores on ordinal categorical items, and for ordinal categorical items, these three types of reliability coefficients represent meanings of reliability with respect to variables' different relations and thus have different values. The model for ordinal categorical items consists of categorical variables Y_{ij}s and latent continuous variables U_{ij}s. Categorical variable Y_{ij} is determined by latent continuous variable U_{ij} and category boundaries C_ks. Some information of U_{ij} is lost when Y_{ij} is given by categorization of U_{ij}. This nonlinear transformation makes the three types of reliability coefficient differ from each other because each coefficient reflects a different aspect of the nonlinear transformation. Precision reliability coefficient ρ_ω of U_{ij}s, which are not discretized and are independent of categorization, has the largest value. Predictive reliability coefficient ρ_{cat}, which is the R^2-index of regression of observed scores Y_is, i.e., sums of discretized values Y_{ij}s of U_{ij}s, on true values F_is, has a smaller value than ρ_ω. Consistency reliability coefficient $\rho_{YY'}$, which represents correlation between parallel tests Y and Y', takes a value between ρ_{cat} and ρ_ω. The popular reliability coefficient α, which is calculated from variances of categorical variables, takes a value similar to $\rho_{YY'}$. Reliability coefficients $\rho_{YY'}$ and α tend to be between ρ_{cat} and ρ_ω (Figures 1 and 2). These tendencies of ordinal categorical items' reliability coefficients were also shown by results from the author's simulations, which are not reported in this study. If the real relationship between observed scores and true values, i.e., intensity of the concept to be measured, is asked, reliability coefficient ρ_{cat}, which denotes the relation of observed scores and true values, should be calculated.

This study proposes an explicit regression model of observed scores on true values (Equation 11). The model of generation of observed scores from true values (Equations 3 and 4) should be discriminated from regression Model 11. Values of Model 11's parameters are determined by the least-squares criterion. The predictive power of Model 11 is given by ρ_{cat}. Since calculation of ρ_{cat} requires some computation time, N_{sub} subsamples were chosen randomly from N_{main} samples of the main MCMC, and reliability coefficients' posterior distributions were calculated for N_{sub} subsamples to reduce computation time. Figure 2 shows that 1,000 samples are sufficient to draw a curve of posterior distributions. Since predictive reliability coefficient ρ_{cat} can clearly differ from other reliability coefficients for items with a small number of categories (cf. Figure 2b), and coefficient ρ_{cat} denotes the relation of observed scores and true values, it is worth calculating.

References

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, 18, 36-52.

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

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

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement*, 27, 440-458.

Fife, D. A., Mendoza, J. L., & Terry, R. (2012). The assessment of reliability under range restriction: A comparison of α, ω, and test-retest reliability for dichotomous data. *Educational and Psychological Measurement*, *72*, 862–888.

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction.* Los Angeles: SAGE Publications.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference, 2nd ed.* New York: Chapman & Hall/CRC.

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

Guilford, J. P. (1954). *Psychometric methods.* New York: McGraw-Hill Book Company, Inc.

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

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. *Psychometrika, 42,* 567–578.

King, B. M., Rosopa, P. J., & Minium, E. W. (2011). *Statistical reasoning in the behavioral sciences.* Hoboken: John Wiley & Sons, Inc.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores with contributions by A. Birnbaum.* Reading: Addison-Wesley Publishing Company.

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.

Raykov, T., Dimitrov, D. M., & Asparouhov, T. (2010). Evaluation of scale reliability with binary measures using latent variable modeling. *Structural Equation Modeling, 17,* 265–279.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory.* New York: Routledge.

Robert, C. P., & Casella, G. (2010a). *Introducing Monte Carlo methods with R.* New York: Springer.

Robert, C. P., & Casella, G. (2010b). *Monte Carlo statistical methods.* New York: Springer.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107–120.

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the

hypothesis of unidimensionality. *Psychometrika, 69,* 613–625.

Yang, Y., & Green, S. B. (2010). A note on structural equation modeling estimates of reliability. *Structural Equation Modeling, 17,* 66–81.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29,* 377–392.

Appendix

The formulae used to calculate ρ_{cat}, $\rho_{YY'}$, and α are listed below:

First, we have

$$\rho_{cat} = \{Cor(Y_i, F_i)\}^2 = \frac{\{Cov(Y_i, F_i)\}^2}{Var(Y_i) \cdot Var(F_i)} = \frac{\{Cov(Y_i, F_i)\}^2}{Var(Y_i)}, \tag{A.1}$$

because, by assumption,

$$\mathrm{Var}(F_i) = 1.$$

We have

$$Var(Y_i) = E\left[\left(Y_i - E(Y_i)\right)^2\right] = E\left(Y_i^2\right) - \{E(Y_i)\}^2. \tag{A.2}$$

By definition of expectation,

$$E(Y_i) = E\left(\sum_{j=1}^{M} Y_{ij}\right) = \sum_{j=1}^{M} E(Y_{ij}) = \sum_{j=1}^{M}\left[\sum_{k=1}^{K} k \cdot P(Y_{ij} = k)\right]. \tag{A.3}$$

We have

$$\begin{aligned} P(Y_{ij} = k) &= P(C_{k-1} \le U_{ij} < C_k) \\ &= \int_{-\infty}^{+\infty} \{\Phi([C_k - (\mu_j + \lambda_j F_i)]/\psi_j) \\ &\quad -\Phi([C_{k-1} - (\mu_j + \lambda_j F_i)]/\psi_j)\}\phi(F_i) dF_i, \end{aligned} \tag{A.4}$$

where $\phi(z)$ is the standard normal distribution function and $\Phi(z)$ is the cumulative distribution function of $\phi(z)$, that is,

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right), \text{ and } \Phi(z) = \int_{-\infty}^{z} \phi(t) \mathrm{dt}.$$

It is assumed that

$$\Phi(-\infty) = 0, \text{ and } \Phi(+\infty) = 1.$$

We have

$$E\left(Y_i^2\right) = E\left[\left(\sum_{j=1}^{M} Y_{ij}\right)^2\right] = \sum_{j=1}^{m} E\left(Y_{ij}^2\right) + \sum_{j1=1}^{m} \sum_{\substack{j2=1 \\ j1 \neq j2}}^{m} E\left(Y_{ij1} Y_{ij2}\right). \tag{A.5}$$

For the terms in Equation A.5, we have

$$E(Y_{ij}^2) = \sum_{k=1}^{K} k^2 \cdot P(Y_{ij} = k) \tag{A.6}$$

and

$$E(Y_{ij1}Y_{ij2}) = \sum_{k1=1}^{K}\sum_{k2=1}^{K} k1 \cdot k2 \cdot P(Y_{ij1} = k1, Y_{ij2} = k2). \tag{A.7}$$

We have

$$\begin{aligned} &P(Y_{ij1} = k1, Y_{ij2} = k2) \\ &= \int_{-\infty}^{+\infty} \{\Phi([C_{k1} - (\mu_{j1} + \lambda_{j1}F_i)]/\psi_{j1}) - \Phi([C_{k1-1} - (\mu_{j1} + \lambda_{j1}F_i)]/\psi_{j1})\} \\ &\times\{\Phi([C_{k2} - (\mu_{j2} + \lambda_{j2}F_i)]/\psi_{j2}) - \Phi([C_{k2-1} - (\mu_{j2} + \lambda_{j2}F_i)]/\psi_{j2})\}\phi(F_i)dF_i. \end{aligned} \tag{A.8}$$

Note that Equation A.8 assumes $j1 \neq j2$.

The numerator of Equation A.1 is

$$\begin{aligned} Cov(Y_i, F_i) &= E[(Y_i - E(Y_i))(F_i - E(F_i))] \\ &= \int_{-\infty}^{+\infty} \left\{ F_i \sum_{j=1}^{M} \left[\sum_{k=1}^{K} k \left(\Phi\left(\left(C_k - (\mu_j + \lambda_j F_i) \right)/\psi_j \right) \right.\right.\right. \\ &\quad \left.\left.\left. - \Phi\left(\left(C_{k-1} - (\mu_j + \lambda_j F_i) \right)/\psi_j \right) \right) \right] \phi(F_i) \right\} dF_i \end{aligned} \tag{A.9}$$

Using Equations A.1 to A.9. ρ_{cat} can be calculated with parameter values μ_js, λ_js, ψ_js and C_ks.

By definition, $\rho_{YY'}$ is given by

$$\rho_{YY'} = Cor(Y_i, Y_i') = \frac{Cov(Y_i, Y_i')}{\sqrt{Var(Y_i)}\sqrt{Var(Y_i')}} = \frac{Cov(Y_i, Y_i')}{Var(Y_i)}, \tag{A.10}$$

where Y_i' is a parallel test of Y_i.

We have

$$Cov(Y_i, Y_i') = \sum_{j1=1}^{M}\sum_{j2=1}^{M} E(Y_{ij1}Y_{ij2}') - \left[E\left(\sum_{j=1}^{M} Y_{ij} \right) \right]^2 \tag{A.11}$$

and

$$E(Y_{ij1}Y'_{ij2}) = \sum_{k1=1}^{K}\sum_{k2=1}^{K} k1 \cdot k2 \cdot P(Y_{ij1} = k1, Y'_{ij2} = k2). \quad (A.12)$$

Probability $P(Y_{ij1} = k1, Y'_{ij2} = k2)$ is given by

$$P(Y_{ij1} = k1, Y'_{ij2} = k2)$$

$$= \int_{-\infty}^{+\infty} \{\Phi([C_{k1} - (\mu_{j1} + \lambda_{j1}F_i)]/\psi_{j1}) - \Phi([C_{k1-1} - (\mu_{j1} + \lambda_{j1}F_i)]/\psi_{j1})\}$$

$$\times\{\Phi([C_{k2} - (\mu_{j2} + \lambda_{j2}F_i)]/\psi_{j2}) - \Phi([C_{k2-1} - (\mu_{j2} + \lambda_{j2}F_i)]/\psi_{j2})\}\phi(F_i)dF_i. \quad (A.13)$$

Note that since Y'_i is a parallel test of Y_i, $j1$ might be equal to $j2$, but in Equation A.8, we assume that $j1 \neq j2$.

$\rho_{YY'}$ can be calculated by Equations A.2, A.3, A.10, A.11, A.12, and A.13.

Coefficient α is given by

$$\alpha = \frac{M}{M-1}\left[1 - \frac{\sum_{j=1}^{M} Var(Y_{ij})}{Var(Y_i)}\right]. \quad (A.14)$$

We have

$$Var(Y_{ij}) = E\left[\{Y_{ij} - E(Y_{ij})\}^2\right] = E(Y_{ij}^2) - [E(Y_{ij})]^2. \quad (A.15)$$

The two terms of Equation A.15 can be calculated as follows:

$$E(Y_{ij}) = \sum_{k=1}^{K} k \cdot P(Y_{ij} = k), \quad (A.16)$$

$$E(Y_{ij}^2) = \sum_{k=1}^{K} k^2 \cdot P(Y_{ij} = k). \quad (A.17)$$

Hence, coefficient α can be calculated by Equations A.2, A.14, A.15, A.16, and A.17.