PSYCHOMETRIKA—VOL. 77, NO. 1, 153–162
JANUARY 2012
DOI: 10.1007/S11336-011-9238-0

A NOTE ON THE RELIABILITY COEFFICIENTS FOR ITEM RESPONSE MODEL-BASED ABILITY ESTIMATES

SEONGHOON KIM

KEIMYUNG UNIVERSITY

Assuming item parameters on a test are known constants, the reliability coefficient for item response theory (IRT) ability estimates is defined for a population of examinees in two different ways: as (a) the product-moment correlation between ability estimates on two parallel forms of a test and (b) the squared correlation between the true abilities and estimates. Due to the bias of IRT ability estimates, the parallel-forms reliability coefficient is not generally equal to the squared-correlation reliability coefficient. It is shown algebraically that the parallel-forms reliability coefficient is expected to be greater than the squared-correlation reliability coefficient, but the difference would be negligible in a practical sense.

Key words: reliability coefficient, ability estimates, item response theory.

1. Introduction

The concept of reliability refers to the precision of test scores and other measurements or the consistency of such measurements when the testing procedure is repeated on a population of individuals or groups (AERA, APA, & NCME, 1985/1999). In the context of classical test theory (CTT), the reliability of observed scores X from a test may be quantified via three indices: the conditional variance of measurement errors E, the average (or marginal) standard error of measurement (SEM), and the reliability coefficient (AERA, APA, & NCME, 1985/1999; Feldt & Brennan, 1989; Feldt, Steffen, & Gupta, 1985; Haertel, 2006). The average SEM is the square root of a weighted average of the conditional error variances across true score T levels and is often simply referred to as the SEM. The SEM (denoted σ_E) is a function of the standard deviation (σ_X) of X and the reliability coefficient ($\rho_{XX'}$) such that $\sigma_E = \sigma_X\sqrt{1-\rho_{XX'}}$ (Lord & Novick, 1968).

In item response theory (IRT), the estimates $\hat{\theta}$ of ability (i.e., latent proficiency) θ parameters for examinees can be viewed as the counterpart of test scores in CTT. This paper primarily concerns the reliability coefficient of estimates of ability based on IRT. Assuming item parameters on a test are known constants (which implies the fit of models to data), the reliability coefficient for IRT ability estimates is generally defined for a population of examinees in two different ways: as (a) the (product-moment) correlation ($\rho_{\hat{\theta}\hat{\theta}'}$) between ability estimates on two parallel forms of a test and (b) the squared correlation ($\rho^2_{\hat{\theta}\theta}$) between the true abilities and estimates. As pointed out by researchers (e.g., Nicewander & Thomasson, 1999; Raju & Oshima, 2005), the reliability coefficient of ability estimates has received less attention than the test information $I(\theta)$ in the IRT area. The reliability coefficient is a global index of precision subject to the ability distribution of examinees (i.e., group-dependent), whereas the test information is conditional on a particular level of ability, providing conditional precision independent of examinee groups (Mellenbergh, 1996). Note that the precision conveyed by the test information is analogous to

Requests for reprints should be sent to Seonghoon Kim, Department of Education, Keimyung University, 1095 Dalgubeoldaero, Dalseo-Gu, Daegu 704-701, South Korea. E-mail: seonghoonkim@empal.com; seonghoonkim@kmu.ac.kr

the reciprocal of the conditional error variance of CTT (AERA, APA, & NCME, 1985/1999; Mellenbergh, 1996). If our primary concern is the conditional precision of IRT ability estimates, therefore, much of our attention to the test information is a natural consequence.

Our interest, however, is often in inter-individual differences on the continuum of the given ability. In this case, our primary concern might be studying the extent to which individual examinees retain their relative positions in a group. The study seeks the information of group-dependent consistency; and thus the reliability coefficient is more relevant than the test information. The reliability coefficient can be useful for other purposes. In general, reliability coefficients (as unit-free summary statistics, ranging from 0 to 1) are most useful in comparing tests or measurement procedures, particularly those that yield scores in different metrics (AERA, APA, & NCME, 1985/1999). This suggests that the reliability coefficient may be efficiently used to compare the psychometric properties of various types of IRT ability estimate, such as the maximum likelihood (ML) estimate (Lord, 1980), the maximum a posteriori (MAP) estimate (Lord 1980, 1986), and the expected a posteriori (EAP) estimate (Bock & Aitkin, 1981; Bock & Mislevy, 1982), and aid in making the best choice for operational use. Such comparison could also be made between the IRT ability estimates and the traditional test scores of CTT, with their respective reliability coefficients. Further, the reliability coefficient can be used to quantify the strength of the (possibly nonlinear) relationship between the ability parameters and estimates. In fact, these practical or potential uses of the reliability coefficient in the IRT area have been illustrated by researchers (Nicewander & Thomasson, 1999; Raju & Oshima, 2005; Samejima, 1994; Sympson, 1980). The potential usefulness is also evidenced by other studies that present formulas for the IRT ability reliability coefficient in the context of computerized adaptive tests (CAT) (Green, Bock, Humphreys, Linn, & Reckase, 1984; Thissen, 1990) or in the context of conventional fixed-length tests (Lord, 1983; Sireci, Thissen, & Wainer, 1991).

However, the present paper was motivated by a finding that the reliability coefficient $\rho_{\hat{\theta}\hat{\theta}'}$ defined by the parallel-forms approach is not generally equal to the reliability coefficient $\rho^2_{\hat{\theta}\theta}$ defined by the squared-correlation approach. This inequality is due to the bias of IRT ability estimates and does not happen for the CTT test scores because they are, by definition, unbiased estimates of true scores (Lord & Novick, 1968). Previous research (Kim & Nicewander, 1993; Lord, 1983; Samejima, 1994; Warm, 1989) showed that this bias occurs for any of the ability estimators developed so far, including the ML, MAP, and EAP estimators. However, as discussed in detail later in this paper, most of the previous studies did not pay much attention to the possible inequality between the $\rho^2_{\hat{\theta}\theta}$ and $\rho_{\hat{\theta}\hat{\theta}'}$, presenting formulas for the IRT coefficient by adopting only one of the two definitions of the reliability coefficient. Furthermore, some formulas, especially for the Bayesian estimates, appear to have been presented under questionable assumptions about the relation between the true θ and the estimates $\hat{\theta}$ (e.g., the conditional expectation of θ on $\hat{\theta}$ is linear). Therefore, the formulas should be regarded as approximations to the exact coefficient.

The present paper has two primary purposes. First, taking into account the bias of ability estimates, this paper presents mathematically rigid and accurate expressions of the $\rho^2_{\hat{\theta}\theta}$ and $\rho_{\hat{\theta}\hat{\theta}'}$ coefficients according to their respective definitions. This makes it possible to diagnose and classify the various formulas that have been presented in previous studies dealing with the topic. Second, this paper provides a general mathematical treatment as to how the two types of coefficient are related to each other and how much they might differ in magnitude. Before performing these tasks, however, this paper first presents the notation and statistical formulas required to efficiently deal with them, and then briefly reviews the formulations of CTT regarding test score reliability.

2. Statistical Notation and Formulas

Throughout, the following statistical notation is used to efficiently denote the moments of variables and the covariance, correlation, and regression between variables. For a variable Y, the mean and variance are denoted by μ_Y and σ_Y^2, respectively. The operator $\mathcal{E}(\cdot)$ is used to express the expected value of a variable, such that $\mathcal{E}(Y) \equiv \mu_Y$. For any two variables Y and Z, their covariance and (product-moment) correlation are denoted by σ_{YZ} and ρ_{YZ}. For the two variables Y and Z, the least-squares linear regression function of Y (dependent variable: DV) on Z (independent variable: IV) is denoted by $R(Y|Z)$. This is a special case of the conditional expectation of Y given Z, $\mathcal{E}(Y|Z)$, which in general may be a nonlinear function of Z. For the linear regression function, the regression coefficient of Y on Z is written by $\beta_{Y|Z} (= \sigma_{YZ}/\sigma_Z^2)$. Lastly, the residual error of predicting Y from Z, $Y - R(Y|Z)$, is denoted by $eR(Y|Z)$.

Also, the following statistical formulas will be used for evaluating variances and covariances when two or more variables are related:

$$\sigma_Y^2 = \sigma_{\mathcal{E}(Y|Z)}^2 + \mathcal{E}_Z\big(\sigma_{Y|Z}^2\big), \tag{1}$$

$$\sigma_Y^2 = \sigma_{R(Y|Z)}^2 + \sigma_{eR(Y|Z)}^2, \tag{2}$$

$$\sigma_{YZ} = \sigma_{\mathcal{E}(Y|Z),Z}, \tag{3}$$

$$\sigma_{Y=Y_1+Y_2,Z=Z_1+Z_2} = \sigma_{Y_1Z_1} + \sigma_{Y_1Z_2} + \sigma_{Y_2Z_1} + \sigma_{Y_2Z_2}, \tag{4}$$

where the subscript Z in the operator $\mathcal{E}_Z(\cdot)$ is used to clearly indicate that the expectation is taken over Z and may be dropped if the context is clear. Equation (1) is known as the ANOVA identity that relates marginal statistics to conditional statistics. Based on this equation, the correlation ratio $\eta_{Y|Z}^2$ for predicting Y from Z is defined as

$$\eta_{Y|Z}^2 = \frac{\sigma_{\mathcal{E}(Y|Z)}^2}{\sigma_Y^2} = 1 - \frac{\mathcal{E}_Z(\sigma_{Y|Z}^2)}{\sigma_Y^2}. \tag{5}$$

Equation (2) states that the variance of a DV is decomposed into a sum of the variance by its linear regression on an IV and the variance of the residuals. The squared correlation (i.e., the coefficient of determination) between Y and Z can be expressed, based on Equation (2), as

$$\rho_{YZ}^2 = \frac{\sigma_{R(Y|Z)}^2}{\sigma_Y^2} = 1 - \frac{\sigma_{eR(Y|Z)}^2}{\sigma_Y^2}. \tag{6}$$

Note that, in general,

$$\rho_{YZ}^2 \le \eta_{Y|Z}^2, \tag{7}$$

because $\sigma_{\mathcal{E}(Y|Z)}^2 \ge \sigma_{R(Y|Z)}^2$ and $\mathcal{E}_Z(\sigma_{Y|Z}^2) \le \sigma_{eR(Y|Z)}^2$ by the relation

$$\sigma_{eR(Y|Z)}^2 = \mathcal{E}_Z\big(\sigma_{Y|Z}^2\big) + \mathcal{E}_Z\big[\big(R(Y|Z) - \mathcal{E}(Y|Z)\big)^2\big].$$

For Equation (7), the equality holds if $\mathcal{E}(Y|Z) = R(Y|Z)$. Equation (3) may be viewed as an extension of Equation (1) to covariance. Equation (4) is often used for computing the covariance between summed variables, based on the component covariances.

3. Review of the Test Score Reliability in Classical Test Theory

In CTT, the observed score X on a test is modeled as a random variable that is the sum of two unobserved components, a true score T and a measurement error E:

$$X = T + E. \tag{8}$$

It should be noted that X is by definition a linearly unbiased estimator of T (Lord & Novick, 1968) so that

$$\mathcal{E}(X|T) = R(X|T) = T. \tag{9}$$

Based on this true-score model, the reliability (coefficient) of a test has been defined in two different ways in CTT (hereafter the two terms "reliability" and "reliability coefficient" will be used interchangeably, unless the context would lead to confusion). First, under the assumption that strictly parallel forms of a test are available, the test reliability has been defined as the correlation $\rho_{XX'}$ between parallel measurements X and $X'(= T + E')$ from the forms (Feldt & Brennan, 1989). By the assumptions associated with the parallel measurements that error scores on a test are uncorrelated with both the true and error scores on different tests, the parallel-forms reliability $\rho_{XX'}$ can be expressed as the ratio of true-score variance to observed-score variance:

$$\rho_{XX'} = \frac{\sigma_{XX'}}{\sigma_X \sigma_{X'}} = \frac{\sigma_{T+E,T+E'}}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_X^2}. \tag{10}$$

Second, the reliability has been defined as the squared correlation ρ_{XT}^2 between observed score and true score (Lord & Novick, 1968, p. 61). This squared correlation can be also expressed as the ratio σ_T^2/σ_X^2:

$$\rho_{XT}^2 = \frac{\sigma_{XT}^2}{\sigma_X^2 \sigma_T^2} = \frac{\sigma_{\mathcal{E}(X|T),T}^2}{\sigma_X^2 \sigma_T^2} = \frac{\sigma_T^2}{\sigma_X^2}. \tag{11}$$

Recognizing that $\rho_{XT}^2 = \eta_{X|T}^2$ by Equation (9), ρ_{XT}^2 may be further expressed as

$$\rho_{XT}^2 = 1 - \frac{\mathcal{E}(\sigma_{X|T}^2)}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}, \tag{12}$$

where $\sigma_E^2 = \mathcal{E}(\sigma_{X|T}^2) = \mathcal{E}(\sigma_{E|T}^2)$.

One can interpret the reliability ρ_{XT}^2 in Equation (11) as the coefficient of determination between the two variables X and T, as regarding X as the DV and T as the IV. Conversely, if one regards T as the DV and X as the IV, the reliability ρ_{XT}^2 can be viewed as the ratio of the variance explained by $R(T|X)$ to true-score variance:

$$\rho_{XT}^2 = \frac{\sigma_{R(T|X)}^2}{\sigma_T^2} = 1 - \frac{\sigma_{eR(T|X)}^2}{\sigma_T^2}, \tag{13}$$

where $R(T|X)$, the so-called Kelley's regressed estimate of T on X, is expressed as

$$R(T|X) = \rho_{XT}^2 X + \left(1 - \rho_{XT}^2\right)\mu_X. \tag{14}$$

4. The Reliability of Ability Estimates in Item Response Theory

Denote by **u** a random vector of an examinee's responses on an n-item test and denote by $\hat{\theta}$ an estimate of the examinee ability parameter θ in IRT. The estimate $\hat{\theta}$ may be any of the ML, MAP, EAP and other legitimate estimates. Consider the estimate as being composed of the ability parameter and an error of estimation (e), such that

$$\hat{\theta} = \theta + e. \tag{15}$$

Based on this modeling of $\hat{\theta}$, which is analogous to Equation (8) for X in CTT, this section will present two definitions of the reliability of $\hat{\theta}$ and show their statistical relation. In the presentation, item parameters on a test are assumed to be known, for example, as a result of previous calibration with a sufficiently large group of examinees.

Before this is done, however, it must be recognized that any IRT ability estimation method so far developed does not provide an unbiased estimator of θ at every possible level of θ (i.e., $-\infty < \theta < +\infty$) unless the test length is infinite. Thus, the conditional expectation of $\hat{\theta}$ for a given θ should be expressed as the sum of the parameter and a bias $B \equiv B(\hat{\theta}|\theta) \equiv \mathcal{E}(e|\theta)$, such that

$$\mathcal{E}(\hat{\theta}|\theta) = \theta + B. \tag{16}$$

The conditional variance of $\hat{\theta}$ on θ, $\sigma^2_{\hat{\theta}|\theta} (= \sigma^2_{e|\theta})$, is called the estimation error variance and its square root is called the standard error of estimate (SEE). Because of the bias of $\hat{\theta}$, the covariance $\sigma_{e\theta}$ in a population is not necessarily equal to zero. This inequality presents a striking contrast to the equality $\sigma_{ET} = 0$ in CTT. For further discussion, note that, by Equations (3) and (4),

$$\sigma_{e\theta} = \sigma_{\mathcal{E}(e|\theta),\theta} = \sigma_{B\theta}; \qquad \sigma_{\hat{\theta}\theta} = \sigma_{\mathcal{E}(\hat{\theta}|\theta),\theta} = \sigma^2_{\theta} + \sigma_{B\theta}; \quad \text{and} \quad \sigma_{\hat{\theta}\theta} = \sigma_{\hat{\theta}(\hat{\theta}-e)} = \sigma^2_{\hat{\theta}} - \sigma_{\hat{\theta}e}.$$

4.1. The Parallel-Forms Reliability

In the context of IRT, any two tests may be viewed as parallel when for each item in one test there is an item in the other test with the same item response function. With this viewpoint, Lord (1983) showed that the parallel-forms reliability coefficient $\rho_{\hat{\theta}\hat{\theta}'}$, the correlation between ability estimates $\hat{\theta}$ and $\hat{\theta}'$ on two parallel tests, can be defined, based on statistical quantities from a single test administration, as

$$\rho_{\hat{\theta}\hat{\theta}'} = \frac{\sigma_{\hat{\theta}\hat{\theta}'}}{\sigma^2_{\hat{\theta}}} = \frac{\sigma^2_{\mathcal{E}(\hat{\theta}|\theta)}}{\sigma^2_{\hat{\theta}}} = \frac{\sigma^2_{\mathcal{E}(\hat{\theta}|\theta)}}{\sigma^2_{\mathcal{E}(\hat{\theta}|\theta)} + \mathcal{E}(\sigma^2_{\hat{\theta}|\theta})} = 1 - \frac{\mathcal{E}(\sigma^2_{\hat{\theta}|\theta})}{\sigma^2_{\hat{\theta}}} = \eta^2_{\hat{\theta}|\theta}. \tag{17}$$

In this equation, because of local independence and parallelism, it follows that

$$\sigma_{\hat{\theta}\hat{\theta}'} \equiv \sigma_{\mathcal{E}(\hat{\theta}|\theta),\mathcal{E}(\hat{\theta}'|\theta)} + \mathcal{E}(\sigma_{\hat{\theta}\hat{\theta}'|\theta}) = \sigma_{\mathcal{E}(\hat{\theta}|\theta),\mathcal{E}(\hat{\theta}'|\theta)} = \sigma^2_{\mathcal{E}(\hat{\theta}|\theta)}.$$

Equation (17) is in form identical to Equation (12) for the CTT reliability. However, it should be noted that $\mathcal{E}(\sigma^2_{X|T}) = \sigma^2_{eR(X|T)}$ in CTT, whereas $\mathcal{E}(\sigma^2_{\hat{\theta}|\theta}) \neq \sigma^2_{eR(\hat{\theta}|\theta)}$ in IRT.

Interestingly, Green et al. (1984) presented Equation (17) as the reliability of $\hat{\theta}$ and referred to it as the marginal reliability. The "marginal" in marginal reliability indicates that the numerator $\mathcal{E}(\sigma^2_{\hat{\theta}|\theta})$ in Equation (17) is an average of the conditional error variance $\sigma^2_{e|\theta} (= \sigma^2_{\hat{\theta}|\theta})$, weighted by the marginal density of θ, say, $g(\theta)$. The marginal (i.e., averaged) property features the error variance σ^2_E in CTT, as shown by Equation (12). Thus, it appears that Green et al. (1984)

presented Equation (17) in IRT as the counterpart of Equation (12) in CTT, without referring to Lord's (1983) derivation for the parallel-forms reliability. On the other hand, Equation (17) for the parallel-forms reliability seems to have been misused for the formula for the squared-correlation reliability. Nicewander and Thomasson (1999), for example, computed the "true" reliability coefficients through simulations based on the third expression in the right-hand side of Equation (17), although they intended to compute the squared-correlation reliability.

4.2. The Squared-Correlation Reliability

In estimation of ability parameters for a population of examinees, a natural concern is how closely related the parameters and the estimates are. This concern leads to defining the reliability of $\hat{\theta}$ as the squared (product-moment) correlation between $\hat{\theta}$ and θ, or

$$\rho^2_{\hat{\theta}\theta} = \frac{\sigma^2_{\hat{\theta}\theta}}{\sigma^2_{\hat{\theta}}\sigma^2_{\theta}}. \tag{18}$$

The correlation coefficient $\rho_{\hat{\theta}\theta}$ is often called the fidelity coefficient in the context of CAT (Sympson, 1980; Weiss, 1982). From the perspective of linear regression analysis, one can interpret $\rho^2_{\hat{\theta}\theta}$ in two different ways. On the one hand, if one regards $\hat{\theta}$ as the DV,

$$\rho^2_{\hat{\theta}\theta} = 1 - \frac{\sigma^2_{eR(\hat{\theta}|\theta)}}{\sigma^2_{\hat{\theta}}} = \frac{\sigma^2_{R(\hat{\theta}|\theta)}}{\sigma^2_{\hat{\theta}}} = \frac{\sigma^2_{\hat{\theta}\theta}/\sigma^2_{\theta}}{\sigma^2_{\hat{\theta}}} = \frac{(\sigma^2_{\theta} + \sigma_{B\theta})^2/\sigma^2_{\theta}}{\sigma^2_{\hat{\theta}}} = \frac{\sigma^2_{\theta}}{\sigma^2_{\hat{\theta}}}(1 + \beta_{e|\theta})^2, \tag{19}$$

where $\beta_{e|\theta}$ is the regression coefficient of e on θ. The linear regression of $\hat{\theta}$ on θ, $R(\hat{\theta}|\theta)$, is expressed as

$$R(\hat{\theta}|\theta) = \beta_{\hat{\theta}|\theta}(\theta - \mu_{\theta}) + \mu_{\hat{\theta}}. \tag{20}$$

If the relations of $\sigma_{\hat{\theta}\theta} = \sigma^2_{\theta}$ and $\mu_{\theta} = \mu_{\hat{\theta}}$ hold as in CTT, the regression is simplified to

$$R(\hat{\theta}|\theta) = \theta. \tag{21}$$

On the other hand, if one regards θ as the DV,

$$\rho^2_{\hat{\theta}\theta} = 1 - \frac{\sigma^2_{eR(\theta|\hat{\theta})}}{\sigma^2_{\theta}} = \frac{\sigma^2_{R(\theta|\hat{\theta})}}{\sigma^2_{\theta}} = \frac{\sigma^2_{\hat{\theta}\theta}/\sigma^2_{\hat{\theta}}}{\sigma^2_{\theta}} = \frac{(\sigma^2_{\hat{\theta}} - \sigma_{\hat{\theta}e})^2/\sigma^2_{\hat{\theta}}}{\sigma^2_{\theta}} = \frac{\sigma^2_{\hat{\theta}}}{\sigma^2_{\theta}}(1 - \beta_{e|\hat{\theta}})^2. \tag{22}$$

In this equation, $R(\theta|\hat{\theta})$ can be viewed as the Kelley's regressed estimate of the true ability θ on $\hat{\theta}$, expressed as

$$R(\theta|\hat{\theta}) = \beta_{\theta|\hat{\theta}}(\hat{\theta} - \mu_{\hat{\theta}}) + \mu_{\theta}. \tag{23}$$

Again, if the relations of $\sigma_{\hat{\theta}\theta} = \sigma^2_{\theta}$ and $\mu_{\theta} = \mu_{\hat{\theta}}$ hold, the regression is expressed similarly to Equation (14) in CTT as follows:

$$R(\theta|\hat{\theta}) = \rho^2_{\hat{\theta}\theta}\hat{\theta} + (1 - \rho^2_{\hat{\theta}\theta})\mu_{\hat{\theta}}. \tag{24}$$

Somewhat surprisingly, the definition of reliability of $\hat{\theta}$ by Equation (19) is not found in the IRT literature, but an expression similar to it is seen in the classic text book by Lord and Novick

(1968, p. 209, Equation (9.8.1)). They defined a generic reliability for observed scores X from a single test as

$$\rho_{XT}^2 = \frac{\sigma_{R(X|T)}^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_X^2}(1+\beta_{E|T})^2,$$

where $\beta_{E|T}$ is the regression coefficient of E (generic errors) on T.

For the reliability expression by Equation (22), it is observed that $\rho_{\hat{\theta}\theta}^2$ has been equalized to $\eta_{\theta|\hat{\theta}}^2$ by the researchers (Nicewander & Thomasson, 1999; Sireci et al., 1991; Sympson, 1980) whose primary interests were in the reliability of Bayesian ability estimates. That is, for the MAP or EAP $\hat{\theta}$, they equalized $R(\theta|\hat{\theta})$ to $\mathcal{E}(\theta|\hat{\theta})$; or equalized the regression residual error variance $\sigma_{eR(\theta|\hat{\theta})}^2$ to a weighted average of the mathematical quantity $\sigma_{\theta|\hat{\theta}}^2$ that they conceived as the Bayesian posterior variance $\sigma_{\theta|\mathbf{u}}^2$ of θ, given a prior distribution for θ, $g(\theta)$, and a response vector $\mathbf{u}$. For example, Sympson (1980) demonstrated, with the argument of $\mathcal{E}(\theta|\hat{\theta}) = R(\theta|\hat{\theta}) = \hat{\theta}$ for the EAP $\hat{\theta}$, that

$$\rho_{\hat{\theta}\theta}^2 = \eta_{\theta|\hat{\theta}}^2 \equiv 1 - \frac{\mathcal{E}(\sigma_{\theta|\hat{\theta}}^2)}{\sigma_\theta^2} \equiv \frac{\sigma_{\mathcal{E}(\theta|\hat{\theta})}^2}{\sigma_\theta^2} = \frac{\sigma_{\hat{\theta}}^2}{\sigma_\theta^2}. \tag{25}$$

Because the relation $\mathcal{E}(\theta|\hat{\theta}) = \hat{\theta}$ implies that $\sigma_{\hat{\theta}}^2 = \sigma_{\mathcal{E}(\theta|\hat{\theta})}^2 = \sigma_\theta^2 - \mathcal{E}(\sigma_{\theta|\hat{\theta}}^2)$, this equation may also be expressed as

$$\rho_{\hat{\theta}\theta}^2 = \eta_{\theta|\hat{\theta}}^2 \equiv \frac{\sigma_{\mathcal{E}(\theta|\hat{\theta})}^2}{\sigma_\theta^2} = \frac{\sigma_{\hat{\theta}}^2}{\sigma_\theta^2} = \frac{\sigma_{\hat{\theta}}^2}{\sigma_{\hat{\theta}}^2 + \mathcal{E}(\sigma_{\theta|\hat{\theta}}^2)}. \tag{26}$$

Although not explicit in the demonstration, Sympson seemed to argue that the relation $\mathcal{E}(\theta|\hat{\theta}) = R(\theta|\hat{\theta}) = \hat{\theta}$ should be implied by the definition of EAP $\hat{\theta} \equiv \mathcal{E}(\theta|\mathbf{u})$. Similarly, Nicewander and Thomasson (1999) presented the squared-correlation reliability coefficient for the MAP (or ML) $\hat{\theta}$ as

$$\rho_{\hat{\theta}\theta}^2 = \eta_{\theta|\hat{\theta}}^2 \equiv 1 - \frac{\mathcal{E}(\sigma_{\theta|\hat{\theta}}^2)}{\sigma_\theta^2} = 1 - \frac{\mathcal{E}(\sigma_{\theta|\mathbf{u}}^2)}{\sigma_\theta^2}. \tag{27}$$

They argued for the relation of $\sigma_{\theta|\hat{\theta}}^2 = \sigma_{\theta|\mathbf{u}}^2$, as follows. If no two items in a test had identical item parameters (in terms of the three-parameter logistic [3PL] model), then there would be a one-to-one mapping of $\mathbf{u}$ into the MAP $\hat{\theta}$. The one-to-one correspondence suggests that the posterior distribution density $h(\theta|\mathbf{u}) = h(\theta|\hat{\theta})$ and the posterior variance $\sigma_{\theta|\mathbf{u}}^2 = \sigma_{\theta|\hat{\theta}}^2$. However, it should be noted that such a one-to-one correspondence may not hold with the one- or two-parameter logistic (1PL or 2PL) model items for the ML $\hat{\theta}$ (see Lord, 1980, p. 57) and for the MAP $\hat{\theta}$, even though the items have different difficulty (b) or discrimination (a) parameters. Although not pointed out by Nicewander and Thomasson (1999), if the one-to-one correspondence holds between $\mathbf{u}$ and the EAP $\hat{\theta}$, Sympson's (1980) argument may be justified because the relation $h(\theta|\mathbf{u}) = h(\theta|\text{EAP}\,\hat{\theta})$ leads to

$$\mathcal{E}(\theta|\mathbf{u}) = \mathcal{E}(\theta|\text{EAP}\,\hat{\theta}) = \text{EAP}\,\hat{\theta}. \tag{28}$$

On the other hand, Sireci et al. (1991) presented formally an expression similar to Equation (27) by designating $\sigma_{\theta|\mathbf{u}}^2$ as $\sigma_{e^*}^2$ and called it the "marginal" reliability to indicate explicitly the marginal feature of $\mathcal{E}(\sigma_{e^*}^2)$. For the presentation, interestingly, they first referred to the marginal reliability definition of Green et al. (1984) as in Equation (17), but later they replaced

$\sigma^2_{\hat{\theta}|\theta}(=\sigma^2_{e|\theta})$ and $\sigma^2_{\hat{\theta}}$ with $\sigma^2_{e^*}$ and σ^2_{θ}, respectively, to present the reliability coefficient for the EAP $\hat{\theta}$. Sireci et al. (1991) stated that the error variance $\sigma^2_{e^*}$ could be computed from (the inverse of) the test information function for the EAP $\hat{\theta}$, and used the computer program MULTILOG (Thissen, 1991) for the computation.

After all, justifiability of Equations (25) to (27) as the exact expressions for the $\rho^2_{\hat{\theta}\theta}$ coefficient of the MAP or EAP $\hat{\theta}$ depends on the tenability of the assumption (or argument) $\mathcal{E}(\theta|\hat{\theta}) = \hat{\theta}$. However, the assumption would be justified only for the EAP $\hat{\theta}$ that have a one-to-one correspondence with every response pattern of $\mathbf{u}$. If the assumption $\mathcal{E}(\theta|\hat{\theta}) = \hat{\theta}$ is not generally true, in fact, $\rho^2_{\hat{\theta}\theta} \leq \eta^2_{\theta|\hat{\theta}}$ and the following inequalities hold among the three expressions for $\rho^2_{\hat{\theta}\theta}$ in Equations (25) to (27): if $\sigma^2_{\mathcal{E}(\theta|\hat{\theta})} \geq \sigma^2_{\hat{\theta}}$,

$$\frac{\sigma^2_{\mathcal{E}(\theta|\hat{\theta})}}{\sigma^2_{\theta}} \geq \frac{\sigma^2_{\hat{\theta}}}{\sigma^2_{\hat{\theta}} + \mathcal{E}(\sigma^2_{\theta|\hat{\theta}})} \geq \frac{\sigma^2_{\hat{\theta}}}{\sigma^2_{\theta}}; \tag{29}$$

otherwise, if $\sigma^2_{\mathcal{E}(\theta|\hat{\theta})} < \sigma^2_{\hat{\theta}}$,

$$\frac{\sigma^2_{\mathcal{E}(\theta|\hat{\theta})}}{\sigma^2_{\theta}} < \frac{\sigma^2_{\hat{\theta}}}{\sigma^2_{\hat{\theta}} + \mathcal{E}(\sigma^2_{\theta|\hat{\theta}})} < \frac{\sigma^2_{\hat{\theta}}}{\sigma^2_{\theta}}. \tag{30}$$

However, little is known regarding the exact relation between $\sigma^2_{\mathcal{E}(\theta|\hat{\theta})}$ and $\sigma^2_{\hat{\theta}}$, partly because IRT does not provide a statistical framework for exactly quantifying the former. Therefore, the three quantities for the Bayesian $\hat{\theta}$, (a) $\eta^2_{\theta|\hat{\theta}}$, (b) $\sigma^2_{\hat{\theta}}/[\sigma^2_{\hat{\theta}} + \mathcal{E}(\sigma^2_{\theta|\hat{\theta}})]$, and (c) $\sigma^2_{\hat{\theta}}/\sigma^2_{\theta}$, all should be regarded as approximate versions of the reliability coefficient $\rho^2_{\hat{\theta}\theta}$. Given the limited knowledge about the relation between $\sigma^2_{\mathcal{E}(\theta|\hat{\theta})}$ and $\sigma^2_{\hat{\theta}}$, it would be wise to use the "intermediate" approximation $\sigma^2_{\hat{\theta}}/[\sigma^2_{\hat{\theta}} + \mathcal{E}(\sigma^2_{\theta|\hat{\theta}})]$ as a basis for assessing the reliability coefficient $\rho^2_{\hat{\theta}\theta}$ for Bayesian ability scores.

4.3. Relations Between $\rho_{\hat{\theta}\hat{\theta}'}$ and $\rho^2_{\hat{\theta}\theta}$

By the definition of true score and the assumption of availability of the parallel forms administered independently, the squared-correlation reliability coefficient (ρ^2_{XT}) is accepted as being equal to the parallel-forms reliability coefficient ($\rho_{XX'}$) in CTT. In contrast, the ability parameter in IRT cannot be simply defined as the expectation of its estimates (i.e., $\hat{\theta}$ is not an unbiased estimator of θ), and thus, in general, the parallel-forms reliability coefficient ($\rho_{\hat{\theta}\hat{\theta}'}$) is not equal to the squared-correlation reliability coefficient ($\rho^2_{\hat{\theta}\theta}$). Analysis of this inconsistency demands knowledge of how the two IRT coefficients $\rho_{\hat{\theta}\hat{\theta}'}$ and $\rho^2_{\hat{\theta}\theta}$ are related to each other and how much they might differ in magnitude.

By re-expression of the correlation coefficient formula, $\rho^2_{\hat{\theta}\theta}$ may be transformed into

$$\rho^2_{\hat{\theta}\theta} = \frac{\sigma^2_{\mathcal{E}(\hat{\theta}|\theta),\theta}}{\sigma^2_{\hat{\theta}}\sigma^2_{\theta}} = \frac{\sigma^2_{\mathcal{E}(\hat{\theta}|\theta)}}{\sigma^2_{\hat{\theta}}} \frac{\sigma^2_{\mathcal{E}(\hat{\theta}|\theta),\theta}}{\sigma^2_{\mathcal{E}(\hat{\theta}|\theta)}\sigma^2_{\theta}} = \rho_{\hat{\theta}\hat{\theta}'}\rho^2_{\mathcal{E}(\hat{\theta}|\theta),\theta}. \tag{31}$$

The squared correlation $\rho^2_{\mathcal{E}(\hat{\theta}|\theta),\theta}$ between $\mathcal{E}(\hat{\theta}|\theta)$ and θ should be less than or equal to 1, so that

$$\rho^2_{\hat{\theta}\theta} \leq \rho_{\hat{\theta}\hat{\theta}'}. \tag{32}$$

Obviously, the equality between $\rho^2_{\hat{\theta}\theta}$ and $\rho_{\hat{\theta}\hat{\theta}'}$ holds when $\mathcal{E}(\hat{\theta}|\theta) = \theta$. However, as noted earlier, the condition is rarely satisfied in the entire range of θ.

The question arises: by how much is $\rho_{\hat{\theta}\hat{\theta}'}$ larger than $\rho^2_{\hat{\theta}\theta}$? To examine this discrepancy, it is useful to express the two coefficients as follows:

$$\rho_{\hat{\theta}\hat{\theta}'} = \frac{\sigma^2_{\mathcal{E}(\hat{\theta}|\theta)}}{\sigma^2_{\hat{\theta}}} = \frac{\sigma^2_{\theta+B}}{\sigma^2_{\hat{\theta}}} = \frac{\sigma^2_{\theta} + 2\sigma_{B\theta} + \sigma^2_{B}}{\sigma^2_{\hat{\theta}}}, \tag{33}$$

$$\rho^2_{\hat{\theta}\theta} = \frac{\sigma^2_{\mathcal{E}(\hat{\theta}|\theta),\theta}}{\sigma^2_{\hat{\theta}}\sigma^2_{\theta}} = \frac{(\sigma^2_{\theta} + \sigma_{B\theta})^2/\sigma^2_{\theta}}{\sigma^2_{\hat{\theta}}} = \frac{\sigma^2_{\theta} + 2\sigma_{B\theta} + (\sigma^2_{B\theta}/\sigma^2_{\theta})}{\sigma^2_{\hat{\theta}}}. \tag{34}$$

It follows that the difference is

$$\delta_{\rho} \equiv \rho_{\hat{\theta}\hat{\theta}'} - \rho^2_{\hat{\theta}\theta} = \frac{\sigma^2_{B} - (\sigma^2_{B\theta}/\sigma^2_{\theta})}{\sigma^2_{\hat{\theta}}}. \tag{35}$$

Recognizing $\sigma^2_B \geq \sigma^2_{B\theta}/\sigma^2_{\theta}$, the difference can be said to be largely dependent on the relative size of σ^2_B compared to $\sigma^2_{\hat{\theta}}$. However, one cannot quantify the difference uniquely because σ^2_B (and $\sigma_{B\theta}$) may vary by the type of $\hat{\theta}$. With the ML $\hat{\theta}$ for an n-item test whose items are analyzed using the 3PL model, Lord (1983) argued that σ^2_B is of order n^{-2} and thus can be neglected in estimation of σ^2_{θ} with a long test. This suggests that the quantity $\sigma^2_{B\theta}/\sigma^2_{\theta}$ is also negligible and, after all, the difference δ_{ρ} will be very small in practice. Therefore, the two coefficients $\rho^2_{\hat{\theta}\theta}$ and $\rho_{\hat{\theta}\hat{\theta}'}$ might be used interchangeably in a practical sense, as long as the number of items is fairly large (e.g., 30 or more).

References

AERA, APA & NCME (1985/1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L., & Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*, 347–360.

Feldt, L.S., & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan.

Feldt, L.S., Steffen, M., & Gupta, N.C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement*, *9*, 351–361.

Haertel, E.H. (2006). Reliability. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education and Praeger.

Kim, J.K., & Nicewander, W.A. (1993). Ability estimation for conventional tests. *Psychometrika*, *58*, 587–599.

Lord, F.M. (1980). *Applications of item response theory to practical testing applications*. Hillsdale, NJ: Erlbaum.

Lord, F.M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*, 233–245.

Lord, F.M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*, 157–162.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mellenbergh, G.J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*, 293–299.

Nicewander, W.A., & Thomasson, G.L. (1999). Some reliability estimates for computerized adaptive tests. *Applied Psychological Measurement*, *23*, 239–247.

Raju, N.S., & Oshima, T.C. (2005). Two prophecy formulas for assessing the reliability of item response theory-based ability estimates. *Educational and Psychological Measurement*, *65*, 361–375.

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

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 237–247.

Sympson, J.B. (1980). *Estimating the reliability of adaptive tests from a single test administration*. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980

Thissen, D. (1990). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 161–186). Hillsdale, NJ: Erlbaum.

Thissen, D. (1991). *MULTILOG: multiple, categorical item analysis and test scoring using item response theory [Computer program]*. Chicago: Scientific Software International.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*, 473–492.

Manuscript Received: 18 SEP 2010
Final Version Received: 14 JUN 2011
Published Online Date: 18 NOV 2011