Understanding Reliability from a Regression Perspective

Yang Liu[1], Jolynn Pek[2], and Alberto Maydeu-Olivares[3]

[1]Department of Human Development and Quantitative Methodology

University of Maryland, College Park

[2]Department of Psychology

The Ohio State University

[3]Department of Psychology

University of South Carolina

Author Note

Correspondence should be made to Yang Liu at 3304R Benjamin Bldg, 3942 Campus Dr, University of Maryland, College Park, MD 20742. Email: yliu87@umd.edu.

Abstract

Reliability is an important quantification of measurement precision based on a latent variable measurement model. Inspired by McDonald (2011), we present a regression framework of reliability, placing emphasis on whether latent or observed scores serve as the regression outcome. Our theory unifies two extant perspectives of reliability: (a) classical test theory (measurement decomposition), and (b) optimal prediction of latent scores (prediction decomposition). Importantly, reliability should be treated as a property of the observed score under a measurement decomposition, but a property of the latent score under a prediction decomposition. To facilitate the evaluation and interpretation of distinct reliability coefficients for complex measurement models, we introduce a Monte Carlo approach for approximate calculation of reliability. We illustrate the proposed computational procedure with an empirical data analysis, which concerns measuring susceptibility and severity of depressive symptoms using a two-dimensional item response theory model. We conclude with a discussion on computing reliability coefficients and outline future avenues of research.

Keywords: reliability, classical test theory, prediction, regression, Monte Carlo methods

Understanding Reliability from a Regression Perspective

Constructs (e.g., personality, intelligence, attitudes), which are not directly observable, are often treated as building blocks in psychological theories (e.g., Cronbach & Meehl, 1955). Often, constructs are operationalized as latent variables (LVs), which are indicated by manifest variables (MVs; also commonly referred to as observed variables, indicator variables, or items). For example, MVs can be participants' responses to items in standardized tests and survey questionnaires, which are designed to measure LVs such as cognitive ability and personality. MVs can also be subtest or test scores.

Observed scores (e.g., summed scores and estimated factor scores) are functions of MVs and are often employed in downstream analysis including scoring, classification, and fitting explanatory models as proxies of latent scores, which are functions of LVs (e.g., see Liu & Pek, 2023). Because observed scores contain measurement error and measurement error might result in biased inferences (Bollen, 1989, Chapter 5; Cole & Preacher, 2014), it is pertinent to quantify this error and adjust our interpretation of inferential results accordingly (e.g., see Schmidt & Hunter, 1999). Measurement error can be assessed by two distinct approaches. One approach is to select an observed score and evaluate the precision with which the observed score can be estimated by LVs. An alternative approach is to select a latent score and to investigate the extent to which it can be predicted by MVs. These two approaches result in the two popular definitions of *reliability*: (a) *classical test theory* (CTT) reliability and (b) *proportional reduction in mean squared error* (PRMSE), respectively.

The contribution of our paper is two-fold. First, we provide a road map for organizing different types of common reliability coefficients, clarifying how they relate to one another, and what distinct information they quantify. Methodological developments in reliability are numerous and nuanced (e.g., Revelle & Condon, 2019), making it challenging for substantive researchers to recognize subtle differences among extant reliability coefficients and employ appropriate coefficients for their research. Thus, our systematization of reliability coefficients, based on

McDonald's (2011) regression perspective, aims to facilitate the proper application of reliability coefficients in substantive research. Drawing on this reliability-as-regression perspective, our second contribution is to introduce a novel and straightforward Monte Carlo (MC) procedure that can be used to approximate reliability coefficients for any measurement model. In linear measurement models (e.g., common factor model; Thurstone, 1935) or in nonlinear measurement models (e.g., IRT; Thissen & Steinberg, 2009) that measure a single LV, reliability estimation is relatively straightforward. We show how our MC procedure provides comparable results to these classical approaches of reliability estimation. More important, we also show that our MC procedure is invaluable when the computation of reliability coefficients becomes intractable in more complex measurement models (e.g., multidimensional IRT models).

The paper is structured as follows. We begin by describing reliability from a regression framework, which stems from McDonald (2011). The framework encompasses two types of decompositions called measurement and prediction decompositions. The measurement decomposition maps onto CTT reliability, and the prediction decomposition maps onto PRMSE. We describe each decomposition using theoretical examples linked to simple measurement models. Next, we introduce an MC procedure for obtaining approximate calculations of reliability coefficients. To illustrate and contrast various types of reliability coefficients, we step through an empirical example on the depressive symptom scale within the Collaborative Psychiatric Epidemiological Surveys (CPES; Magnus & Liu, 2022), making use of the novel MC procedure for computing reliability coefficients. Finally, we discuss implications of our work on reporting reliability coefficients in substantive research and outline future directions of research.

Reliability From a Regression Framework

Assumptions and Notations

We will restrict our discussion of reliability at the level of the population, assuming that the specified measurement model is (or at least closely approximates) the data generating mechanism and that all the parameters in the measurement model are known. In practice,

however, parameters in the measurement model must be estimated from a finite sample of MVs. The model can be either linear (e.g,. factor analysis) or nonlinear (e.g., IRT). Let $\mathbf{y}_i$ be the $m \times 1$ vector of MVs associated with person i, and $\boldsymbol{\eta}_i$ be the $d \times 1$ vector of LVs for person i. Here, m denotes the number of MVs (i.e., the length of the measurement instrument) and d denotes the number of LVs (i.e., dimensionality of the measurement model). To distinguish between random variables (i.e., variables before they are observed) and their empirical realizations (i.e., variables when they take on fixed values), we underline the random variables.

For ease of presenting theoretical examples and without loss of generality, we focus on unidimensional scores. For person i, let $s(\mathbf{y}_i)$ be an *observed score* (a function of MVs $\mathbf{y}_i$) and $\xi(\boldsymbol{\eta}_i)$ be a *latent score* (a function of LVs $\boldsymbol{\eta}_i$). Both observed and latent scores are unidimensional as implied by non-bolded expressions of $s(\mathbf{y})$ and $\xi(\boldsymbol{\eta})$, respectively. Observed scores are often used to rank or classify individuals as well as get incorporated in downstream statistical analyses. An observed score $s(\underline{\mathbf{y}}_i)$ is a function of MVs $\underline{\mathbf{y}}_i$ and depends stochastically on LVs $\boldsymbol{\eta}_i$ through the conditional distribution $s(\underline{\mathbf{y}}_i)|\boldsymbol{\eta}_i$. A latent score $\xi(\underline{\boldsymbol{\eta}}_i)$ can never be observed but can be inferred from the conditional distribution of $\xi(\underline{\boldsymbol{\eta}}_i)|\mathbf{y}_i$. Both conditional distributions can be obtained from the joint distribution of $\underline{\mathbf{y}}_i$ and $\underline{\boldsymbol{\eta}}_i$, with probability density function (pdf) $f(\mathbf{y}_i, \boldsymbol{\eta}_i)$. Stated differently, these conditional distributions are fully determined by the measurement model.

Below, we describe two simple measurement models at the level of the population to illustrate different types of observed and latent scores. After describing these examples, we introduce our regression framework of reliability which encompass CTT reliability and PRMSE.

Examples and Observed Scores

Example 1: One-Factor Model

Consider a common factor model with a single LV $\underline{\eta}_i$ ($d = 1$), whose mean equals to 0 and variance equals to φ. The m MVs in $\underline{\mathbf{y}}_i$ are associated with the LV by the following equation:

$$\underline{\mathbf{y}}_i = \boldsymbol{\nu} + \boldsymbol{\lambda}\underline{\eta}_i + \underline{\boldsymbol{\epsilon}}_i, \tag{1}$$

in which $\boldsymbol{\nu}$ is an $m \times 1$ vector of MV intercepts and $\boldsymbol{\lambda}$ is an $m \times 1$ vector of factor loadings. The $m \times 1$ vector of unique factors, denoted by $\underline{\boldsymbol{\epsilon}}_i$, are uncorrelated with $\underline{\eta}_i$. The covariance among the unique factors, $\text{Cov}(\underline{\boldsymbol{\epsilon}}_i, \underline{\boldsymbol{\epsilon}}_i)$ is denoted by $\boldsymbol{\Theta}$, which is an $m \times m$ diagonal matrix. As an example, consider a one-factor model for three MVs ($m = 3$) with parameters

$$\boldsymbol{\nu} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \ \boldsymbol{\lambda} = \begin{bmatrix} 0.3 \\ 0.5 \\ 0.7 \end{bmatrix}, \ \varphi = 1, \text{ and } \boldsymbol{\Theta} = \begin{bmatrix} 0.91 & 0 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0 & 0.51 \end{bmatrix}. \tag{2}$$

Equation 2 implies that the LV $\underline{\eta}_i$ and the three MVs in $\underline{\mathbf{y}}_i = [y_{i1}, y_{i2}, y_{i3}]'$ are standardized.

Due to linearity of the one-factor model, commonly used observed scores are weighted or unweighted sums of MVs. Let us first consider the regression factor score (Thomson, 1936; Thurstone, 1935),

$$s_1(\mathbf{y}_i) = \frac{\boldsymbol{\lambda}'\boldsymbol{\Theta}^{-1}\mathbf{y}_i}{\boldsymbol{\lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\lambda} + \varphi^{-1}}, \tag{3}$$

which is a weighted sum of the elements in $\mathbf{y}_i$. Let us further assume that the LV follows a normal distribution, $\underline{\eta}_i \sim \mathcal{N}(0, \varphi)$, and that the unique factors follow (independent to the LV) a normal distribution as well, $\underline{\boldsymbol{\epsilon}}_i \sim \mathcal{N}(\mathbf{0}_m, \boldsymbol{\Theta})$, in which $\mathbf{0}_m$ denotes an $m \times 1$ vector of zeros. Taken together, the regression factor score $s_1(\mathbf{y}_i)$ coincides with the conditional mean of $\underline{\eta}_i$ given $\mathbf{y}_i$, $\mathbb{E}(\underline{\eta_i}|\mathbf{y}_i)$. The conditional mean is also known as the expected *a posteriori* [EAP] score of the LV $\underline{\eta}_i$ (Thissen & Thissen-Roe, 2022). As such, we use the term "regression factor scores" interchangeably with "EAP scores" in the linear one-factor model example. The second observed score under consideration is the (unweighted) summed score of the m MVs,

$$s_2(\mathbf{y}_i) = \mathbf{1}_m'\mathbf{y}_i, \tag{4}$$

in which $\mathbf{1}_m$ is an $m \times 1$ vector of ones. Given the model parameter values in Equation 2, the regression factor score for individual i is $s_1(\mathbf{y}_i) = 0.14y_{i1} + 0.28y_{i2} + 0.57y_{i3}$ (by Equation 3) and the summed score is $s_2(\mathbf{y}_i) = y_{i1} + y_{i2} + y_{i3}$ (by Equation 4).

Under a unidimensional measurement model, the latent scores of interest can be the LV

itself; i.e., $\xi_1(\eta_i) = \eta_i$. Alternatively, it might be desirable to align the scale of the latent score to the scale of an observed score (e.g., the summed score). For instance, let the expected summed score be

$$\xi_2(\eta_i) = \mathbb{E}\left[s_2(\underline{\mathbf{y}}_i)|\eta_i\right] = \mathbf{1}_m'(\boldsymbol{\nu} + \boldsymbol{\lambda}\eta_i), \tag{5}$$

in which $\mathbb{E}[\cdot|\eta_i]$ is the conditional expectation over the MVs given the LV η_i. Because $\xi_2(\eta_i)$ is a linear transformation of η_i, $\xi_2(\eta_i)$ is perfectly correlated with η_i. Substituting $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$ with population parameter values from Equation 2 results in Equation 5 becoming $\xi_2(\eta_i) = 1.5\eta_i$.

Example 2: Two-Parameter Logistic Model

In our second example, the MVs are dichotomous item responses. Conditional on a single LV η_i, assume that the MVs $\underline{y}_{ij}$, $j = 1, \ldots, m$, are independent (i.e., local independence; McDonald, 1994; Stout, 2002). The conditional probability of $\underline{y}_{ij} = k$ given η_i, where $k \in \{0, 1\}$, is parameterized by a two-parameter logistic (2PL) model (Birnbaum, 1968):

$$f_j(k|\eta_i) = \mathbb{P}\{\underline{y}_{ij} = k|\eta_i\} = \frac{\exp\left[k(\alpha_j + \beta_j\eta_i)\right]}{1 + \exp(\alpha_j + \beta_j\eta_i)}, \tag{6}$$

in which α_j and β_j are the intercept and slope parameters for the jth MV, respectively. Equation 6 is commonly referred to as the item response function (IRF). We further assume that the LV $\underline{\eta}_i$ follows a standard normal distribution: $\underline{\eta}_i \sim \mathcal{N}(0, 1)$. As a numerical example, consider a test composed of $m = 3$ dichotomous MVs with

$$\boldsymbol{\alpha} = \begin{bmatrix} 1 \\ 0 \\ -2 \end{bmatrix}, \text{ and } \boldsymbol{\beta} = \begin{bmatrix} 1 \\ 1.5 \\ 2 \end{bmatrix}, \tag{7}$$

in which $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ contain the intercept and slope parameters, respectively.

Similar to the linear one-factor model example, we consider two types of observed scores. The first type of score is the EAP score of $\underline{\eta}_i$ based on the 2PL model:

$$s_1(\mathbf{y}_i) = \mathbb{E}(\underline{\eta}_i|\mathbf{y}_i) = \frac{\int \eta_i f(\mathbf{y}_i|\eta_i)\phi(\eta_i)d\eta_i}{f(\mathbf{y}_i)}, \tag{8}$$

Table 1

Marginal probabilities (Equation 10), expected a posteriori (EAP) scores of η_i (Equation 8), and summed scores for all eight response patterns in the illustrative example with three two-parameter logistic items (see Equations 6 and 7). The R package `mirt` *with the default tuning setup was used to compute EAP scores.*

Pattern	Probability	EAP score	Summed score
$[0, 0, 0]'$	0.19	−0.96	0
$[1, 0, 0]'$	0.26	−0.41	1
$[0, 1, 0]'$	0.08	−0.16	1
$[0, 0, 1]'$	0.01	0.08	1
$[1, 1, 0]'$	0.24	0.31	2
$[1, 0, 1]'$	0.04	0.54	2
$[0, 1, 1]'$	0.02	0.76	2
$[1, 1, 1]'$	0.15	1.22	3

in which

$$f(\mathbf{y}_i|\eta_i) = \prod_{j=1}^{m} f_j(y_{ij}|\eta_i) \tag{9}$$

denotes the conditional pdf of $\underline{\mathbf{y}}_i$ given η_i and $\phi(\eta)$ denotes the standard normal pdf of $\underline{\eta}_i$. Additionally,

$$f(\mathbf{y}_i) = \int f(\mathbf{y}_i|\eta_i)\phi(\eta_i)d\eta_i \tag{10}$$

denotes the marginal pdf of $\underline{\mathbf{y}}_i$. The integrals in Equations 8 and 10 do not have closed-form expressions and are often approximated by numerical quadrature in practice.[1] With three binary MVs, there are $2^3 = 8$ response patterns. Using item parameter values specified in Equation 7, we computed the marginal probabilities and EAP scores for these 8 response patterns, which are summarized in Table 1. The second type of score is the summed score $s_2(\mathbf{y}_i) = \mathbf{1}'_m\mathbf{y}_i$ (presented in the fourth column of Table 1). Note that there are only four distinct values for the summed score for the eight response patterns in a three-item test (see Table 1).

In addition to latent scores that represent the LV, $\xi_1(\eta_i) = \eta_i$, the expected summed score (also known as the test characteristic curve; e.g., Thissen & Wainer, 2001, pp. 159-160) is another

[1] Consistent with the default configuration in the R package `mirt` (Chalmers, 2012), we always use 61 equally-spaced quadrature nodes ranging from −6 to 6 to approximate intractable integrals in the 2PL example.

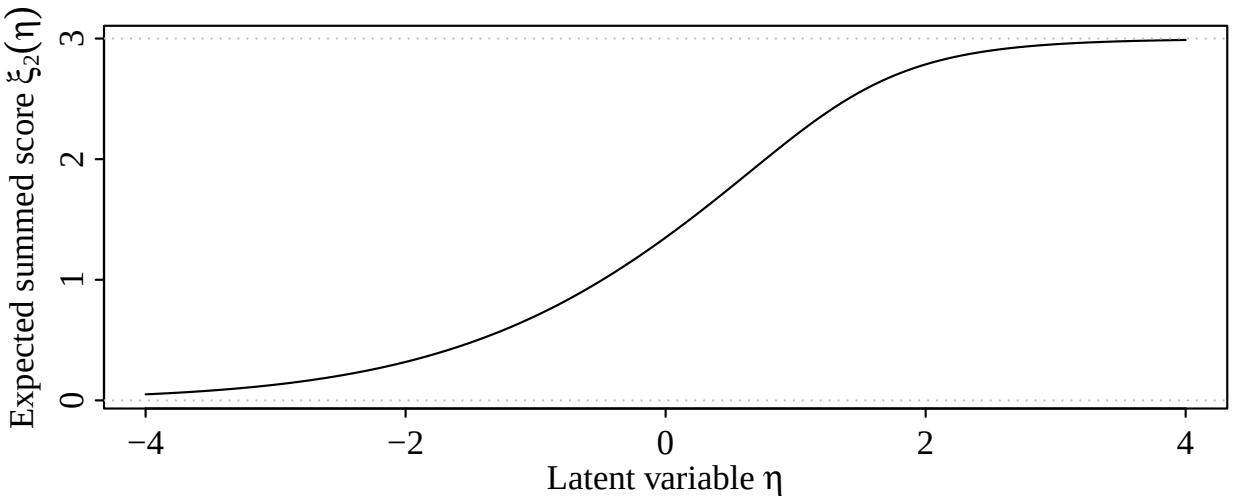

Figure 1

Expected summed score (i.e., Equation 11) as a nonlinear, strictly monotone function of latent variable. Item parameters used to generate the plot can be found in Equation 7. Horizontal dotted lines are shown at 0 and 3, which are the lower and upper asymptotes of the curve.

popular latent score in IRT applications. The expected summed score is expressed as

$$\xi_2(\eta_i) = \mathbb{E}\left[s_2(\underline{\mathbf{y}}_i)|\eta_i\right] = \sum_{j=1}^{m} f_j(1|\eta_i). \tag{11}$$

In contrast to Equation 5 for the linear one-factor model example, Equation 11 is often a nonlinear function of the LV, η. Figure 1 is a visualization of Equation 11 using values of the slope and intercept parameters specified in Equation 7. When all the slope parameters are positive, $\xi_2(\eta_i)$ is a nonlinear and strictly increasing function of η_i, which is bounded between 0 and the total number of MVs (i.e., $m = 3$ in the current example).

In the next section, we review two approaches to defining reliability coefficients, mirroring the measurement and prediction decompositions by McDonald (2011). Both decompositions stem from regression equations but differ in whether a latent score or an observed score serves as the outcome variable. The measurement decomposition yields the well-known definition of reliability coefficients in CTT, while the prediction decomposition introduces PRMSE as a measure of reliability that is more frequently seen in the IRT literature.

Measurement Decomposition

Classical Test Theory

We adapt classical definitions and results in Lord and Novick (1968, p. 34) to the context of LV measurement models using our notation to describe the measurement decomposition of reliability. Recall that an underlying LV model determines the joint distribution of MVs $\underline{\mathbf{y}}_i$ and LVs $\underline{\boldsymbol{\eta}}_i$. It follows that the conditional distribution of an observed score $s(\underline{\mathbf{y}}_i)$ given the LVs $\boldsymbol{\eta}_i$, denoted by $s(\underline{\mathbf{y}}_i)|\boldsymbol{\eta}_i$, is uniquely identified for every $\boldsymbol{\eta}_i$. This conditional distribution is called the *propensity distribution* of the observed score $s(\underline{\mathbf{y}}_i)$ (cf. Lord & Novick, 1968, p. 29–38), which reflects the variability of observed scores across independent and identically distributed (i.i.d.) measurement instances for the same person i.[2] The *true score* corresponding to $s(\underline{\mathbf{y}}_i)$ is defined as the expectation of the propensity distribution; i.e., $\mathbb{E}\big[s(\underline{\mathbf{y}}_i)|\boldsymbol{\eta}_i\big]$ (cf. Lord & Novick, 1968, Definition 2.3.1). By definition, the true score is the long-run average of observed scores across repeated i.i.d. measurements of the same person, which can be conceived as a version of the observed score that is free of measurement error. The error score for person i, ε_i (cf. Lord & Novick, 1968, Definition 2.4.1.), is defined as the difference between the observed score and the true score

$$\varepsilon_i = s(\mathbf{y}_i) - \mathbb{E}\big[s(\underline{\mathbf{y}}_i)|\boldsymbol{\eta}_i\big]. \tag{12}$$

Over the population of persons, the error score has a mean of zero and is uncorrelated with the true score. That is, $\mathbb{E}\,\underline{\varepsilon}_i = 0$ and $\mathrm{Cov}\big(\underline{\varepsilon}_i, \mathbb{E}\big[s(\underline{\mathbf{y}}_i)|\underline{\boldsymbol{\eta}}_i\big]\big) = 0$. Note that $\boldsymbol{\eta}_i$ is underlined in $\mathbb{E}\big[s(\underline{\mathbf{y}}_i)|\underline{\boldsymbol{\eta}}_i\big]$ in the covariance expression because the conditional expectation $\mathbb{E}\big[s(\underline{\mathbf{y}}_i)|\underline{\boldsymbol{\eta}}_i\big]$ is a function of $\underline{\boldsymbol{\eta}}_i$ and $\underline{\boldsymbol{\eta}}_i$ is treated as random. CTT reliability of $s(\underline{\mathbf{y}}_i)$, denoted as $\mathrm{Rel}_{\mathrm{CTT}}(s(\underline{\mathbf{y}}_i))$, can be defined in two main ways described below.

First, CTT reliability quantifies the proportion of observed score variance that can be

[2] In a different view, the conditional distribution $s(\underline{\mathbf{y}}_i)|\boldsymbol{\eta}_i$ captures the variability of observed scores across persons with the same LV value. Correspondingly, the conditional expectation $\mathbb{E}[s(\underline{\mathbf{y}}_i)|\boldsymbol{\eta}_i]$ is the average of observed scores within the subpopulation at the LV level $\boldsymbol{\eta}_i$. See Holland (1990) for relevant discussions.

explained by true scores. Mathematically, CTT reliability is the ratio of true score variance over observed score variance, or one minus the ratio of error score variance over observed score variance:

$$\mathrm{Rel}_{\mathrm{CTT}}\big(s(\underline{\mathbf{y}}_i)\big) = \frac{\mathrm{Var}\big(\mathbb{E}\big[s(\underline{\mathbf{y}}_i)|\underline{\boldsymbol{\eta}}_i\big]\big)}{\mathrm{Var}\big[s(\underline{\mathbf{y}}_i)\big]} = 1 - \frac{\mathrm{Var}(\underline{\varepsilon}_i)}{\mathrm{Var}\big[s(\underline{\mathbf{y}}_i)\big]}. \tag{13}$$

The second equality in Equation 13 follows from the fact that the true score is uncorrelated with the error score. Second, assuming that the true score variance $\mathrm{Var}\big(\mathbb{E}\big[s(\underline{\mathbf{y}}_i)|\underline{\boldsymbol{\eta}}_i\big]\big)$ is positive, CTT reliability also equals to the squared correlation between the observed score $s(\underline{\mathbf{y}}_i)$ and the true score $\mathbb{E}\big[s(\underline{\mathbf{y}}_i)|\underline{\boldsymbol{\eta}}_i\big]$:

$$\mathrm{Rel}_{\mathrm{CTT}}\big(s(\underline{\mathbf{y}}_i)\big) = \mathrm{Corr}\big(s(\underline{\mathbf{y}}_i), \mathbb{E}\big[s(\underline{\mathbf{y}}_i)|\underline{\boldsymbol{\eta}}_i\big]\big)^2. \tag{14}$$

In sum, CTT reliability can be interpreted in two exchangeable ways: a ratio of true- and observed-score variances (Equation 13) and a squared-correlation between true and observed scores (Equation 14).[3]

A Regression Formulation

Rearranging Equation 12 leads to the widely known *true score formula*:

$$s(\mathbf{y}_i) = \mathbb{E}\big[s(\underline{\mathbf{y}}_i)|\boldsymbol{\eta}_i\big] + \varepsilon_i, \tag{15}$$

or equivalently in words,

observed score = true score + measurement error.

---

[3] A third definition of CTT reliability can be defined using the concept of a parallel score, which is less relevant to the present paper. Let $\tilde{s}_i$ be another observable score produced by person i. Then, $\tilde{s}_i$ is *parallel* to $s(\mathbf{y}_i)$ if (a) $\mathbb{E}(\underline{\tilde{s}}_i|\boldsymbol{\eta}_i) = \mathbb{E}\big[s(\underline{\mathbf{y}}_i)|\boldsymbol{\eta}_i\big]$ (i.e., the true scores for $\tilde{s}$ and $s(\mathbf{y}_i)$ are equal), (b) $\mathrm{Var}(\underline{\varepsilon}_i) = \mathrm{Var}(\underline{\tilde{\varepsilon}}_i)$ (i.e., the variance of error scores are equal), where $\tilde{\varepsilon}_i = \tilde{s}_i - \mathbb{E}(\underline{\tilde{s}}_i|\boldsymbol{\eta}_i)$ denotes the error score of $\tilde{s}_i$, and (c) $\mathrm{Cov}(\underline{\varepsilon}_i, \underline{\tilde{\varepsilon}}_i) = 0$ (i.e., the error scores are uncorrelated; Lord & Novick, 1968, p. 47). Thus, CTT reliability equals to the correlation between two parallel test scores (cf. Lord & Novick, 1968, p. 58): $\mathrm{Rel}_{\mathrm{CTT}}\big(s(\underline{\mathbf{y}}_i)\big) = \mathrm{Corr}\big[s(\underline{\mathbf{y}}_i), \underline{\tilde{s}}_i\big]$. The parallel score $\tilde{s}_i$, often regarded as a concept, might not be necessarily computed from MVs. A more intuitive approach to construct parallel scores, which requires stronger assumptions, was described by Lord (1983; see also Kim, 2012). Let $\tilde{\mathbf{y}}_i$ denote an i.i.d. copy of the MV vector $\mathbf{y}_i$ conditional on $\boldsymbol{\eta}_i$; we may think of $\tilde{\mathbf{y}}_i$ as independent responses to an equivalent form of the same measurement instrument produced by the same person. Lord (1983) directly set $\tilde{s}_i = s(\tilde{\mathbf{y}}_i)$, which satisfies all the three requirements of a parallel score; the independence requirement, however, is stronger than having uncorrelated error scores.

Equation 15 is hereinafter termed the *measurement decomposition* of reliability (McDonald, 2011). Recall that a regression traces the conditional expectation of an outcome variable given one or more explanatory variables (e.g., Fox, 2015, p. 15). Thus, Equation 15 can be viewed as the regression of the observed score $s(\underline{\mathbf{y}}_i)$ on all the LVs $\underline{\boldsymbol{\eta}}_i$. This regression is in general nonlinear because the predicted value of the model, which coincides with the true score $\mathbb{E}\left[s(\underline{\mathbf{y}}_i)|\boldsymbol{\eta}_i\right]$ might not be a linear function of $\boldsymbol{\eta}_i$ (e.g., in a 2PL model). If we regard the true score as the explanatory variable, then Equation 15 can be alternatively understood as a *unit-weight linear regression* (i.e., with an intercept of zero and a slope of one) of the observed score $s(\underline{\mathbf{y}}_i)$ on the true score $\mathbb{E}\left[s(\underline{\mathbf{y}}_i)|\underline{\boldsymbol{\eta}}_i\right]$. This second regression interpretation follows from the towering property of the conditional expectation: $\mathbb{E}\left\{s(\underline{\mathbf{y}}_i)|\mathbb{E}\left[s(\underline{\mathbf{y}}_i)|\boldsymbol{\eta}_i\right]\right\} = \mathbb{E}\left\{\mathbb{E}\left[s(\underline{\mathbf{y}}_i)|\boldsymbol{\eta}_i\right]|\mathbb{E}\left[s(\underline{\mathbf{y}}_i)|\boldsymbol{\eta}_i\right]\right\} = \mathbb{E}\left[s(\underline{\mathbf{y}}_i)|\boldsymbol{\eta}_i\right]$. Thus, when the observed score $s(\underline{\mathbf{y}}_i)$ is regressed onto the true score $\mathbb{E}\left[s(\underline{\mathbf{y}}_i)|\boldsymbol{\eta}_i\right]$, the predicted value is simply the true score itself, leading to a unit-weight regression.

Let $\varrho^2(\underline{u}, \underline{\mathbf{v}})$ denote the population coefficient of determination when regressing an outcome variable $\underline{u}$ on explanatory variables $\underline{\mathbf{v}}$: The regression is implied by the joint distribution of $\underline{u}$ and $\underline{\mathbf{v}}$ and is potentially nonlinear. Both views of Equation 15 (i.e., regression of the observed score on all the LVs and regression of the observed score on the true score) yield the same coefficient of determination that equals to CTT reliability (Equation 13):

$$\varrho^2\left(s(\underline{\mathbf{y}}_i), \underline{\boldsymbol{\eta}}_i\right) = \varrho^2\left(s(\underline{\mathbf{y}}_i), \mathbb{E}\left[s(\underline{\mathbf{y}}_i)|\underline{\boldsymbol{\eta}}_i\right]\right) = \mathrm{Rel}_{\mathrm{CTT}}\left(s(\underline{\mathbf{y}}_i)\right).$$

In words, reliability from the measurement decomposition perspective is the amount of variance in the observed score $s(\underline{\mathbf{y}}_i)$ explained by the latent score $\underline{\boldsymbol{\eta}}_i$ (i.e., treating $\underline{\boldsymbol{\eta}}_i$ as explanatory variables) or by the true score $\mathbb{E}\left[s(\underline{\mathbf{y}}_i)|\underline{\boldsymbol{\eta}}_i\right]$ (i.e., treating $\mathbb{E}\left[s(\underline{\mathbf{y}}_i)|\underline{\boldsymbol{\eta}}_i\right]$ as the explanatory variable). Implicit in this interpretation is the assumption that the chosen observed score (which is a summary statistic of the MVs) reflects underlying LVs (constructs). Stated differently, the LVs give rise to the MVs that are eventually combined to form an observed score (cf. Borsboom & Mellenbergh, 2002).

Below, we illustrate how to calculate CTT reliability from the regression framework for

the examples of a one-factor model (Equation 1 with parameter values in Equation 2) and a 2PL model (Equation 6 with parameter values Equation 7). Within each model, CTT reliability coefficients are computed for the EAP score $s_1(\underline{\mathbf{y}}_i)$ and the summed score $s_2(\underline{\mathbf{y}}_i)$.

Example 1: One-Factor Model

Consider first the regression factor score, which is also the EAP score under the normality assumption (i.e., $s_1(\mathbf{y}_i)$; Equation 3). By Equations 1 and 3, the true score of $s_1(\underline{\mathbf{y}}_i)$ is

$$\mathbb{E}\left[s_1(\underline{\mathbf{y}}_i)|\eta_i\right] = \frac{\boldsymbol{\lambda}'\boldsymbol{\Theta}^{-1}(\boldsymbol{\nu} + \boldsymbol{\lambda}\eta_i)}{\boldsymbol{\lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\lambda} + \varphi^{-1}}. \tag{16}$$

That is, the regression of the EAP score $s_1(\underline{\mathbf{y}}_i)$ on the LV $\underline{\eta}_i$ is linear. With covariance algebra, it can be shown that the observed score variance of the EAP score $s_1(\underline{\mathbf{y}}_i)$ is

$$\mathrm{Var}\left[s_1(\underline{\mathbf{y}}_i)\right] = \frac{\varphi\boldsymbol{\lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\lambda}}{\boldsymbol{\lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\lambda} + \varphi^{-1}}, \tag{17}$$

and the corresponding true score variance (over the population distribution of $\underline{\eta}_i$) is

$$\mathrm{Var}\left(\mathbb{E}\left[s_1(\underline{\mathbf{y}}_i)|\underline{\eta}_i\right]\right) = \varphi\left[\frac{\boldsymbol{\lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\lambda}}{\boldsymbol{\lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\lambda} + \varphi^{-1}}\right]^2. \tag{18}$$

Taking the ratio of Equation 18 over Equation 17 yields CTT reliability for the EAP score:

$$\varrho^2\left(s_1(\underline{\mathbf{y}}_i), \underline{\boldsymbol{\eta}}_i\right) = \frac{\mathrm{Var}\left(\mathbb{E}\left[s_1(\underline{\mathbf{y}}_i)|\underline{\eta}_i\right]\right)}{\mathrm{Var}\left[s_1(\underline{\mathbf{y}}_i)\right]} = \frac{\boldsymbol{\lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\lambda}}{\boldsymbol{\lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\lambda} + \varphi^{-1}}. \tag{19}$$

Because the true score underlying the EAP score $s_1(\underline{\mathbf{y}}_i)$ is a linear function of the LV $\underline{\eta}_i$ (see Equation 16), Equation 19 is not only the squared correlation between $s_1(\underline{\mathbf{y}}_i)$ and its true score (i.e., the squared-correlation interpretation of CTT reliability) but also the squared correlation between $s_1(\underline{\mathbf{y}}_i)$ and $\underline{\eta}_i$. When evaluated at the model parameter values in Equation 2, CTT reliability for the regression factor score is 0.58.

The regression of the summed score $s_2(\underline{\mathbf{y}}_i)$ on the latent score $\underline{\eta}_i$ is also linear. Observe that the true score underlying $s_2(\underline{\mathbf{y}}_i)$, given by Equation 5, is a linear function of η_i. The observed

score variance and the true score variance associated with summed score $s_2(\underline{\mathbf{y}}_i)$ are

$$\text{Var}\left[s_2(\underline{\mathbf{y}}_i)\right] = \varphi(\mathbf{1}_m'\boldsymbol{\lambda})^2 + \text{tr}(\boldsymbol{\Theta}), \tag{20}$$

and

$$\text{Var}\left(\mathbb{E}\left[s_2(\underline{\mathbf{y}}_i)|\underline{\eta}_i\right]\right) = \varphi(\mathbf{1}_m'\boldsymbol{\lambda})^2, \tag{21}$$

respectively. CTT reliability for the summed score, expressed as the ratio of Equation 21 over Equation 20 is

$$\varrho^2\left(s_2(\underline{\mathbf{y}}_i), \underline{\eta}_i\right) = \frac{\text{Var}\left(\mathbb{E}\left[s_2(\underline{\mathbf{y}}_i)|\underline{\eta}_i\right]\right)}{\text{Var}\left[s_2(\underline{\mathbf{y}}_i)\right]} = \frac{\varphi(\mathbf{1}_m'\boldsymbol{\lambda})^2}{\varphi(\mathbf{1}_m'\boldsymbol{\lambda})^2 + \text{tr}(\boldsymbol{\Theta})}. \tag{22}$$

Equation 22 is widely known as the coefficient omega (McDonald, 1999, p. 89). Due to the linearity of the regression, Equation 22 gives not only the squared correlation between the summed score and its underlying true score but also the squared correlation between the summed score and the LV. With values specified for φ, $\boldsymbol{\lambda}$, and $\boldsymbol{\Theta}$ in Equation 2, CTT reliability for the summed score (Equation 22) is equal to 0.51, which is noticeably lower than that of the EAP score (i.e., 0.58).

Example 2: Two-Parameter Logistic Model

Reliability calculations are typically less tractable for nonlinear measurement models. As will be shown, calculations for the simple 2PL example become rather involved. Thus, we introduce a more intuitive and widely applicable numerical recipe for approximate reliability calculation for nonlinear measurement models in the section “A Monte Carlo Procedure."

Consider the EAP score of η_1, denoted $s_1(\mathbf{y}_i)$. The true score underlying $s_1(\mathbf{y}_i)$ is

$$\mathbb{E}\left[s_1(\underline{\mathbf{y}}_i)|\eta_i\right] = \sum_{\mathbf{y}_i} s_1(\mathbf{y}_i) f(\mathbf{y}_i|\eta_i), \tag{23}$$

in which the summation is taken over all 2^m possible response patterns. The regression of the EAP score $s_1(\underline{\mathbf{y}}_i)$ on the LV $\underline{\eta}_i$ is nonlinear due to the logistic IRFs on the right-hand side of Equation 23. We proceed by computing the error score and observed score variances. The error

score variance can be expressed as

$$\mathrm{Var}(\underline{\varepsilon}_i) = \mathbb{E}\big(\mathrm{Var}\big[s_1(\underline{\mathbf{y}}_i)|\underline{\eta}_i\big]\big) = \int \mathrm{Var}\big[s_1(\underline{\mathbf{y}}_i)|\eta_i\big]\phi(\eta_i)d\eta_i, \tag{24}$$

in which the conditional variance of the EAP score $s_1(\underline{\mathbf{y}}_i)$ given the LV η_i is

$$\mathrm{Var}\big[s_1(\underline{\mathbf{y}}_i)|\eta_i\big] = \sum_{\mathbf{y}_i} s_1(\mathbf{y}_i)^2 f(\mathbf{y}_i|\eta_i) - \left[\sum_{\mathbf{y}_i} s_1(\mathbf{y}_i) f(\mathbf{y}_i|\eta_i)\right]^2. \tag{25}$$

Meanwhile, the observed variance of the EAP score $s_1(\underline{\mathbf{y}}_i)$ can be computed as

$$\mathrm{Var}\big[s_1(\underline{\mathbf{y}}_i)\big] = \sum_{\mathbf{y}_i} s_1(\mathbf{y}_i)^2 f(\mathbf{y}_i) - \left[\sum_{\mathbf{y}_i} s_1(\mathbf{y}_i) f(\mathbf{y}_i)\right]^2. \tag{26}$$

CTT reliability for the EAP score $s_1(\underline{\mathbf{y}}_i)$ is then equal to one minus the ratio of Equation 24 over Equation 26:

$$\varrho^2\big(s_1(\underline{\mathbf{y}}_i), \underline{\eta}_i\big) = 1 - \frac{\mathbb{E}\big(\mathrm{Var}\big[s_1(\underline{\mathbf{y}}_i)|\underline{\eta}_i\big]\big)}{\mathrm{Var}\big[s_1(\underline{\mathbf{y}}_i)\big]} = 1 - \frac{\int \mathrm{Var}\big[s_1(\underline{\mathbf{y}}_i)|\eta_i\big]\phi(\eta_i)d\eta_i}{\sum_{\mathbf{y}_i} s_1(\mathbf{y}_i)^2 f(\mathbf{y}_i) - \big[\sum_{\mathbf{y}_i} s_1(\mathbf{y}_i) f(\mathbf{y}_i)\big]^2}. \tag{27}$$

In practice, Equation 27 can only be evaluated approximately as integrals involved in Equation 24 and the marginal pdf of the MVs $f(\mathbf{y}_i)$ defined by Equation 10, and are computationally intractable. Using the item parameters in Equation 7 and numerical quadrature to approximate integrals, we obtain 0.51 as CTT reliability of the EAP score $s_1(\underline{\mathbf{y}}_i)$ in the three-item example. When the EAP score $s_1(\underline{\mathbf{y}}_i)$ is the observed score of interest, its underlying true score (Equation 23) is a nonlinear function of the LV η_i. Hence, the resulting CTT reliability (Equation 27), while remaining to be the squared correlation between the observed score and its underlying true score, is no longer the same as the squared correlation between the observed score and the LV. As we will see in the "Prediction Decomposition" section, the squared correlation between the observed score and the LV coincides with PRMSE of the LV $\underline{\eta}_i$.

CTT reliability for the summed score $s_2(\underline{\mathbf{y}}_i)$ can be calculated in the same fashion.

$$\varrho^2\big(s_2(\underline{\mathbf{y}}_i), \underline{\eta}_i\big) = 1 - \frac{\mathbb{E}\big(\mathrm{Var}\big[s_2(\underline{\mathbf{y}}_i)|\underline{\eta}_i\big]\big)}{\mathrm{Var}\big[s_2(\underline{\mathbf{y}}_i)\big]} = 1 - \frac{\int \mathrm{Var}\big[s_2(\underline{\mathbf{y}}_i)|\eta_i\big]\phi(\eta_i)d\eta_i}{\sum_{\mathbf{y}_i} s_2(\mathbf{y}_i)^2 f(\mathbf{y}_i) - \big[\sum_{\mathbf{y}_i} s_2(\mathbf{y}_i) f(\mathbf{y}_i)\big]^2}. \tag{28}$$

The conditional variance of the summed score (cf. Equation 25 for the EAP score) has the following simpler expression:

$$\mathrm{Var}\big[s_2(\underline{\mathbf{y}}_i)|\eta_i\big] = \sum_{j=1}^{m} f_j(0|\eta_i) f_j(1|\eta_i),$$

which is derived from assuming local independence and using the variance formula for a Bernoulli random variable. In contrast to the one-factor example, CTT reliability of the summed score $s_2(\underline{\mathbf{y}}_i)$ is different from the squared correlation between the summed score and the LV $\underline{\eta}_i$, for the reason that the true score underlying the summed score (Equation 11) is a nonlinear transformation of the LV. Under the 2PL model with item parameters in Equation 7, CTT reliability of the summed score is 0.5, which is only slightly lower than that of the EAP score (i.e., 0.51).

Prediction Decomposition

Optimal Prediction of a Latent Score

The prediction decomposition of reliability concerns using MVs to predict a latent score, $\xi(\underline{\boldsymbol{\eta}}_i)$, which is a function of the LVs $\underline{\boldsymbol{\eta}}_i$. The uncertainty in prediction can be quantified by the mean squared error (MSE). Generally speaking, a predictor is (a possibly constant) function of MVs and is optimal when it minimizes the MSE. Let us first consider making predictions without MVs (i.e., using only a constant predictor $g \in \mathbb{R}$). A well-known result in elementary statistics is that the optimal predicting constant of the latent score $\xi(\underline{\boldsymbol{\eta}}_i)$ is its unconditional mean $\mathbb{E}\,\xi(\underline{\boldsymbol{\eta}}_i)$ (e.g., Casella & Berger, 2002, Example 2.2.6), and the minimized MSE is

$$\min_{g\in\mathbb{R}} \mathbb{E}\big[\xi(\underline{\boldsymbol{\eta}}_i) - g\big]^2 = \mathbb{E}\big[\xi(\underline{\boldsymbol{\eta}}_i) - \mathbb{E}\,\xi(\underline{\boldsymbol{\eta}}_i)\big]^2 = \mathrm{Var}\big[\xi(\underline{\boldsymbol{\eta}}_i)\big]. \tag{29}$$

Now suppose that we employ the MVs $\underline{\mathbf{y}}_i$ to predict the latent score $\xi(\underline{\boldsymbol{\eta}}_i)$ by considering

predictors of the form $g(\underline{\mathbf{y}}_i)$, where g denotes a real-valued function of the MVs. Similar to Equation 29, it can be shown that the MSE-minimizing predictor is $\mathbb{E}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i\right]$, i.e., the EAP score of $\xi(\underline{\boldsymbol{\eta}}_i)$ (e.g., Casella & Berger, 2002, Exercise 4.13). The minimized MSE then equals to

$$\min_g \mathbb{E}\left[\xi(\underline{\boldsymbol{\eta}}_i) - g(\underline{\mathbf{y}}_i)\right]^2 = \mathbb{E}\left\{\xi(\underline{\boldsymbol{\eta}}_i) - \mathbb{E}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i\right]\right\}^2 = \mathbb{E}\left(\mathrm{Var}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i\right]\right). \tag{30}$$

When predicting $\xi(\underline{\boldsymbol{\eta}}_i)$ with a constant, the minimized MSE (Equation 29) coincides with the unconditional variance of the latent score, $\mathrm{Var}\left[\xi(\underline{\boldsymbol{\eta}}_i)\right]$. This unconditional variance can be interpreted as the total amount of variability (uncertainty) due to individual differences in $\xi(\underline{\boldsymbol{\eta}}_i)$. When the MVs $\underline{\mathbf{y}}_i$ are used for predicting the latent score $\xi(\underline{\boldsymbol{\eta}}_i)$, the minimized MSE, which can be understood as the remaining uncertainty, is reduced to $\mathbb{E}\left(\mathrm{Var}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i\right]\right)$. The amount of uncertainty reduction, expressed as $\mathrm{Var}\left[\xi(\underline{\boldsymbol{\eta}}_i)\right] - \mathbb{E}\left(\mathrm{Var}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i\right]\right) = \mathrm{Var}\left(\mathbb{E}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i\right]\right)$, thus reflects the degree to which we can precisely recover the latent score using MVs. From this perspective, measurement error can be understood as the remaining uncertainty in the latent score $\xi(\underline{\boldsymbol{\eta}}_i)$ that cannot be captured by the MVs $\mathbf{y}_i$. When the remaining uncertainty in prediction is zero, the latent score can be reproduced from the MVs without error. Taken together, define

$$\mathrm{PRMSE}\left(\xi(\underline{\boldsymbol{\eta}}_i)\right) = 1 - \frac{\mathbb{E}\left(\mathrm{Var}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i\right]\right)}{\mathrm{Var}\left[\xi(\underline{\boldsymbol{\eta}}_i)\right]} = \frac{\mathrm{Var}\left(\mathbb{E}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i\right]\right)}{\mathrm{Var}\left[\xi(\underline{\boldsymbol{\eta}}_i)\right]}. \tag{31}$$

PRMSE quantifies the proportion of latent score variance that can be explained by the MVs. Assuming that the EAP score has a positive variance, $\mathrm{Var}\left(\mathbb{E}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i\right]\right) > 0$, it can also be shown that $\mathrm{PRMSE}\left(\xi(\underline{\boldsymbol{\eta}}_i)\right) = \mathrm{Corr}\left(\xi(\underline{\boldsymbol{\eta}}_i), \mathbb{E}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i\right]\right)^2$ (e.g., Kim, 2012); that is, PRMSE equals to the squared correlation between the latent score $\xi(\underline{\boldsymbol{\eta}}_i)$ and its EAP score $\mathbb{E}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i\right]$. Note further that PRMSE for predicting $\xi(\underline{\boldsymbol{\eta}}_i)$ cannot exceed the CTT reliability of the EAP score $\mathbb{E}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i\right]$, which follows from Equation 31 of Kim (2012, p.160).

A Regression Formulation

PRMSE can also be approached from a regression perspective. In contrast to the measurement decomposition in which an observed score serves as the regression outcome

(Equation 15), the *prediction decomposition* treats the latent score as the outcome (McDonald, 2011) in

$$\xi(\boldsymbol{\eta}_i) = \mathbb{E}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\mathbf{y}_i\right] + \delta_i, \tag{32}$$

or equivalently in words,

$$\text{latent score} = \text{EAP score} + \text{prediction error}.$$

The prediction error δ_i is defined as the difference between the latent score $\xi(\boldsymbol{\eta}_i)$ and its EAP score $\mathbb{E}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\mathbf{y}_i\right]$. Equation 32 implies that the prediction error has mean zero (i.e., $\mathbb{E}\underline{\delta}_i = 0$) and is uncorrelated with the EAP score (i.e., $\text{Cov}(\underline{\delta}_i, \mathbb{E}[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i]) = 0$) in the population; these properties are analogous to the facts that the error score has mean zero (i.e., $\mathbb{E}\underline{\varepsilon}_i = 0$) and is uncorrelated with the true score (i.e., $\text{Cov}(\underline{\varepsilon}_i, \mathbb{E}[s(\underline{\mathbf{y}}_i)|\underline{\boldsymbol{\eta}}_i]) = 0$). The orthogonality of prediction error with the EAP score allows Equation 32 to be interpreted as a regression of the latent score $\xi(\underline{\boldsymbol{\eta}}_i)$ onto all the MVs $\underline{\mathbf{y}}_i$. Treating the EAP score as the explanatory variable in the prediction decomposition, Equation 32 can be alternatively viewed as a unit-weight linear regression of the latent score $\xi(\underline{\boldsymbol{\eta}}_i)$ on the EAP score $\mathbb{E}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i\right]$.

The prediction decomposition of the latent score $\xi(\underline{\boldsymbol{\eta}}_i)$, whether viewed as a potentially nonlinear regression on $\underline{\mathbf{y}}_i$ or a unit-weight linear regression on the EAP score, $\mathbb{E}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i\right]$, yields a coefficient of determination that equals to the PRMSE:

$$\varrho^2\left(\xi(\underline{\boldsymbol{\eta}}_i), \underline{\mathbf{y}}_i\right) = \varrho^2\left(\xi(\underline{\boldsymbol{\eta}}_i), \mathbb{E}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i\right]\right) = \text{PRMSE}\left(\xi(\underline{\boldsymbol{\eta}}_i)\right).$$

In words, reliability from the prediction decomposition is interpreted as the amount of variance in the latent score $\xi(\underline{\boldsymbol{\eta}}_i)$ explained by the MVs $\underline{\mathbf{y}}_i$ or equivalently the EAP estimator of the latent score, denoted $\mathbb{E}\left[\xi(\underline{\boldsymbol{\eta}}_i)|\underline{\mathbf{y}}_i\right]$. Different from the measurement decomposition, the prediction decomposition does not emphasize the reflective nature of the measurement model (i.e., MVs are manifestations of underlying LVs).

Illustrations on calculating the PRMSE reliability coefficient for the one-factor and 2PL

models are presented in the next section. For each model, we calculate PRMSE for the LV $\xi_1(\eta_i) = \eta_i$ and the expected summed score $\xi_2(\eta_i)$ (Equation 11).

Example 1: One-Factor Model

Assuming that the common factor $\underline{\eta}_i$ and the unique factors $\underline{\boldsymbol{\epsilon}}_i$ (Equation 1) are normally distributed, the EAP estimate of $\xi_1(\eta_i) = \eta_i$ is identical to the regression factor score $s_1(\mathbf{y}_i)$ (Thissen & Thissen-Roe, 2022). The expression of $s_1(\mathbf{y}_i)$ under a one-factor model (i.e., Equation 3) implies that the regression of the LV $\underline{\eta}_i$ on the MVs $\underline{\mathbf{y}}_i$ is linear. Note that the variance of the regression factor score $\text{Var}[s_1(\underline{\mathbf{y}}_i)]$ has been derived in Equation 17 and the variance of the latent score is $\text{Var}[\xi_1(\underline{\eta}_i)] = \text{Var}(\underline{\eta}_i) = \varphi$. Thus, the PRMSE of $\xi_1(\underline{\eta}_i)$ is the ratio of these two variances:

$$\varrho^2\big(\xi_1(\underline{\eta}_i), \underline{\mathbf{y}}_i\big) = \frac{\text{Var}\big[s_1(\mathbf{y}_i)\big]}{\text{Var}\big[\xi_1(\underline{\eta}_i)\big]} = \frac{\text{Var}\big(\mathbb{E}\big[s_1(\underline{\mathbf{y}}_i)|\underline{\eta}_i\big]\big)}{\text{Var}\big[s_1(\underline{\mathbf{y}}_i)\big]} = \frac{\boldsymbol{\lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\lambda}}{\boldsymbol{\lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\lambda} + \varphi^{-1}}. \tag{33}$$

Note that PRMSE for $\xi_1(\underline{\eta}_i) = \underline{\eta}_i$ under a linear normal one-factor model (Equation 33) is equivalent to CTT reliability for the regression factor score $s_1(\underline{\mathbf{y}}_i)$ ((Equation 19). This is to be expected because the regression of $s_1(\underline{\mathbf{y}}_i)$ on the LV $\underline{\eta}_i$ is linear (expressed by Equation 16) and the regression of the LV $\underline{\eta}_i$ on $s_1(\underline{\mathbf{y}}_i)$ is also linear (with unit weight; see Equation 32). Taken together, the two coefficients of determination from the measurement and prediction decompositions coincide and are equal to the squared Pearson correlation between the regression factor score $s_1(\underline{\mathbf{y}}_i)$ and the LV $\underline{\eta}_i$.[4] With specific model parameter values in Equation 2, PRMSE of the latent score $\xi_1(\underline{\eta}_i) = \underline{\eta}_i$ is 0.58 and identical to CTT reliability of the regression factor score $s_1(\underline{\mathbf{y}}_i)$.

Recall that the expected summed score $\xi_2(\eta_i)$ for the linear one-factor model is also a linear transformation of the LV η_i (Equation 5). Thus, $\text{PRMSE}\big(\xi_2(\underline{\eta}_i)\big) = \text{PRMSE}\big(\xi_1(\underline{\eta}_i)\big) = 0.58$ for this numerical example. Note that PRMSE of $\xi_2(\underline{\eta}_i)$ is usually distinct from (and higher than) CTT reliability of the summed score $s_2(\underline{\mathbf{y}}_i)$ (Equation 22; cf. PRMSE of $\xi_1(\underline{\eta}_i)$ is identical to CTT reliability of $s_1(\underline{\mathbf{y}}_i)$). To see why, recall that the summed score $s_2(\underline{\mathbf{y}}_i)$ and the expected

[4] An alternative explanation is that $s_1(\underline{\mathbf{y}}_i)$ and $\underline{\eta}_i$ follow a bivariate normal distribution.

summed score $\xi_2(\underline{\eta}_i)$ follow a bivariate normal distribution jointly. Regressing $s_2(\underline{\mathbf{y}}_i)$ on $\xi_2(\underline{\eta}_i)$ and regressing $\xi_2(\underline{\eta}_i)$ on $s_2(\underline{\mathbf{y}}_i)$ result in the same coefficient of determination. While the formal coefficient of determination is CTT reliability of $s_2(\underline{\eta}_i)$, the latter coefficient is in general smaller than PRMSE of $\xi_2(\underline{\eta}_i)$. This is because the summed score $s_2(\mathbf{y}_i)$ is a summary statistic of the MVs $\mathbf{y}_i$ (i.e., summed scores have less information than $\mathbf{y}_i$), and a regression onto the summed score $s_2(\underline{\mathbf{y}}_i)$ usually produces a smaller coefficient of determination than a regression onto the MVs $\underline{\mathbf{y}}_i$.

Example 2: Two-Parameter Logistic Model

For the first latent score that is the LV itself (i.e., $\xi_1(\underline{\eta}_i) = \underline{\eta}_i$), its EAP estimator is $s_1(\underline{\mathbf{y}}_i)$, which we have defined in Equation 8. The variance of $s_1(\underline{\mathbf{y}}_i)$ has been expressed in Equation 26, and the variance of the LV η_i is one. As such, PRMSE of the first latent score $\xi_1(\eta_i)$ is

$$\varrho^2\big(\xi_1(\underline{\eta}_i), \mathbf{y}_i\big) = \frac{\operatorname{Var}\big[s_1(\underline{\eta}_i)|\underline{\mathbf{y}}_i\big]}{\operatorname{Var}(\underline{\eta}_i)} = \sum_{\mathbf{y}_i} s_1(\mathbf{y}_i)^2 f(\mathbf{y}_i) - \left[\sum_{\mathbf{y}_i} s_1(\mathbf{y}_i) f(\mathbf{y}_i)\right]^2. \tag{34}$$

Table 1 presents the response-pattern probabilities of the MVs (i.e., $f(\mathbf{y}_i)$) and the EAP estimates of $\xi_1(\eta_i)$ (i.e., $s_1(\mathbf{y}_i)$) for the three-item 2PL example. Plugging in the specified values of the population parameters (Equation 7) into Equation 34, we have $\operatorname{PRMSE}\big(\xi_1(\underline{\eta}_i)\big) = 0.5$.

In the prediction decomposition (Equation 32) of the expected summed score $\xi_2(\underline{\eta}_i)$ (Equation 11), the predicted value is its EAP estimate:

$$s_3(\mathbf{y}_i) = \mathbb{E}\big[\xi_2(\underline{\eta}_i)|\mathbf{y}_i\big] = \frac{\int \left[\sum_{j=1}^{m} f_j(1|\eta_i)\right] f(\mathbf{y}_i|\eta_i)\phi(\eta_i)d\eta_i}{f(\mathbf{y}_i)}, \tag{35}$$

which we momentarily denote by $s_3(\mathbf{y}_i)$ for ease of reference. Similar to Equation 26, the variance of $s_3(\underline{\mathbf{y}}_i)$ over the population distribution of $\underline{\mathbf{y}}_i$ is given by

$$\operatorname{Var}\big[s_3(\underline{\mathbf{y}}_i)\big] = \sum_{\mathbf{y}_i} s_3(\mathbf{y}_i)^2 f(\mathbf{y}_i) - \left[\sum_{\mathbf{y}_i} s_3(\mathbf{y}_i) f(\mathbf{y}_i)\right]^2. \tag{36}$$

In addition, the variance of the expected summed score $\xi_2(\underline{\eta}_i)$ can be calculated as

$$\mathrm{Var}\left[\xi_2(\underline{\eta}_i)\right] = \int \xi_2(\eta_i)^2 \phi(\eta_i) d\eta_i - \left[\int \xi_2(\eta_i)\phi(\eta_i) d\eta_i\right]^2, \tag{37}$$

which does not possess a closed-form expression. PRMSE for the expected summed score $\xi_2(\underline{\eta}_i)$ is then the ratio of Equation 36 to Equation 37:

$$\varrho^2\left(\xi_2(\underline{\eta}_i), \underline{\mathbf{y}}_i\right) = \frac{\mathrm{Var}\left[s_3(\underline{\mathbf{y}}_i)\right]}{\mathrm{Var}\left[\xi_2(\underline{\boldsymbol{\eta}}_i)\right]} = \frac{\sum_{\mathbf{y}_i} s_3(\mathbf{y}_i)^2 f(\mathbf{y}_i) - \left[\sum_{\mathbf{y}_i} s_3(\mathbf{y}_i) f(\mathbf{y}_i)\right]^2}{\int \xi_2(\eta_i)^2 \phi(\eta_i) d\eta_i - \left[\int \xi_2(\eta_i)\phi(\eta_i) d\eta_i\right]^2}. \tag{38}$$

For the 2PL example, we first compute Equation 35 (i.e., $s_3(\mathbf{y}_i)$, the EAP estimates of expected summed scores) for all eight response patterns as ordered in Table 1:

$$(0.80, 1.11, 1.28, 1.45, 1.63, 1.80, 1.97, 2.29)'.$$

Then Equation 36 (i.e., the variance of $s_3(\underline{\mathbf{y}}_i)$) and Equation 37 (i.e., the variance of $\xi_2(\underline{\eta}_i)$) are estimated as 0.24 and 0.46, respectively. Plugging these values into Equation 38, we obtain PRMSE of the expected summed score $\xi_2(\underline{\eta}_i)$ as 0.52.

A Monte Carlo Procedure

Estimating Reliability by Simulation

In principle, reliability coefficients compatible with the regression framework can be directly evaluated as functions of model parameters (as we have done for the two simple illustrative examples). When the measurement model is complex (e.g., a multidimensional IRT model), however, computational difficulties in evaluating expectations over the spaces of LVs and MVs ensue. For example, expectations with respect to LVs can be intractable and numerical integration can become inefficient when the latent dimensionality is high. In addition, expectations with respect to discrete MVs often translate to a finite sum across all possible response patterns whose number grows exponentially as the number of MVs increases. As a workaround to such computational challenges, reliability coefficients can be approximated by simulation.

Step 1. *Simulate latent scores and MVs.* Given a measurement model with known parameters

(e.g., using estimated parameters from data), simulate a large sample of i.i.d. LV vectors $\boldsymbol{\eta}_i$, $i = 1, \ldots, M$, in which M is the Monte Carlo sample size. Then for each i, simulate an MV vector $\mathbf{y}_i$ conditional on $\boldsymbol{\eta}_i$.

Step 2. *Nonparametric regression.* For a measurement decomposition, compute the observed score $s(\mathbf{y}_i)$ for each $i = 1, \ldots, M$ and fit a nonparametric regression predicting the observed scores by all the LVs. For a prediction decomposition, compute the latent score $\xi(\boldsymbol{\eta}_i)$ for each $i = 1, \ldots, M$ and then fit a nonparametric regression predicting the latent scores by all the MVs.

Step 2′. *Unit weight regression.* For a measurement decomposition, compute the observed score $s(\mathbf{y}_i)$ and its true score $\mathbb{E}\left[s(\mathbf{y}_i)|\boldsymbol{\eta}_i\right]$ for each $i = 1, \ldots, M$, followed by fitting a unit-weight regression predicting the observed scores by the true scores. For a prediction decomposition, compute the latent score $\xi(\boldsymbol{\eta}_i)$ and its EAP estimate $\mathbb{E}\left[\xi(\boldsymbol{\eta}_i)|\mathbf{y}_i\right]$ for each $i = 1, \ldots, M$ and then fit a unit-weight regression predicting the latent scores by the EAP estimates.

Step 3. *Estimate reliability coefficient*. Obtain the sample coefficient of determination (e.g., the R^2 statistic) from the fitted regression, which is an estimate of reliability.

Step 1 produces a large sample of LVs and MVs based on a specified measurement model, which can be utilized to approximate the joint distribution of MVs $\underline{\mathbf{y}}_i$ (and thus observed scores $s(\underline{\mathbf{y}}_i)$) and LVs $\underline{\boldsymbol{\eta}}_i$ (and thus latent scores $\xi(\underline{\boldsymbol{\eta}}_i)$). Step 2 and Step 2′ involve fitting regressions to the MC samples. In practice, only one of the two steps needs to be performed. Step 2 is generally applicable when a nonparametric regression can be estimated with high precision. As an alternative, Step 2′ is useful when the computation of true scores or EAP estimates is viable, avoiding the fitting of nonparametric regression. The explanatory variables in the regression may be discrete; for example, in measurement decompositions with discrete LVs (e.g., latent profile analysis) and prediction decompositions with discrete MVs (e.g., binary item responses). In cases with discrete MVs, Step 2 reduces to fitting a linear regression to dummy-coded patterns for

Table 2
True values and Monte Carlo (MC) estimates (R^2) of reliability coefficients for the linear one-factor example (Equations 1 and 2). MC algorithms with Step 2 versus Step 2′ lead to R^2 statistics that coincide up to the fourth digit. LV: Latent variable.

Coefficient	Decomposition	Regression outcome	True reliability	R^2
$\varrho^2(s_1(\underline{\mathbf{y}}_i), \underline{\eta}_i)$ (Equation 19)	measurement	regression factor score	0.5821	0.5825
$\varrho^2(s_2(\underline{\mathbf{y}}_i), \underline{\eta}_i)$ (Equation 22)	measurement	summed score	0.5090	0.5091
$\varrho^2(\xi_1(\underline{\boldsymbol{\eta}}_i), \underline{\mathbf{y}}_i)$ (Equation 33)	prediction	LV	0.5821	0.5825

Table 3
True values and Monte Carlo (MC) estimates (R^2) of reliability coefficients under the 2PL example (Equations 6 and 7). MC algorithms with Step 2 versus Step 2′ lead to R^2 statistics that coincide up to the fourth digit. LV: Latent variable. EAP: Expected a posteriori.

Coefficient	Decomposition	Regression outcome	True reliability	R^2
$\varrho^2(s_1(\underline{\mathbf{y}}_i), \underline{\eta}_i)$ (Equation 27)	measurement	EAP score of LV	0.5146	0.5137
$\varrho^2(s_2(\underline{\mathbf{y}}_i), \underline{\eta}_i)$ (Equation 28)	measurement	summed score	0.4951	0.4942
$\varrho^2(\xi_1(\underline{\boldsymbol{\eta}}_i), \underline{\mathbf{y}}_i)$ (Equation 34)	prediction	LV	0.4960	0.4953
$\varrho^2(\xi_2(\underline{\boldsymbol{\eta}}_i), \underline{\mathbf{y}}_i)$ (Equation 38)	prediction	expected summed score	0.5150	0.5141

explanatory variables.

Throughout the rest of the paper, all MC-based reliability estimates are computed from $M = 10^6$ MC samples in Step 1, which ensures negligible Monte Carlo error. In Step 2, the R package `mgcv` (Wood, 2017) was used to fit nonparametric regressions and the default thin-plate spline smoother (Wood, 2003) was adopted. R code for numerical illustrations in the present paper will be provided as Supplemental Material when the paper is accepted for publication.

Tables 2 and 3 present MC-based reliability estimates for the linear one-factor model and the 2PL model, respectively. MC estimates (i.e., R^2) are juxtaposed against their population values that have been presented in earlier sections. Observe that the maximum absolute deviation between the MC-based estimates and population values is less than 0.001, demonstrating accuracy of approximation for these examples. Moreover, Step 2 (i.e., nonparametric regression) and Step 2′ (unit weight regression) yield estimates that coincide up to the fourth digit.

Empirical Example

The two simple examples described above were designed for two conceptual purposes. First, these examples illustrated the distinction between the measurement and prediction decompositions of reliability, emphasizing the regression framework. Second, these examples demonstrated the accuracy of the novel MC algorithm. When the true parameter values are known, the MC-based reliability estimates closely approximated their population counterparts (see Tables 2 and 3). Next, we apply the MC algorithm to an empirical example to illustrate calculating reliability coefficients from the measurement and prediction decompositions in a more realistic and complex measurement model.

Data and Measurement Model

The data are from the National Comorbility Survey Replication (NCS-R), which is part of the Collaborative Psychiatric Epidemiological Surveys (CPES; Alegria, Jackson, Kessler, & Takeuchi, 2001-2003). We focus on a 14-item scale that measures depressive symptoms in the past 30 days. Each item was rated on a scale of 0 = "never," 1 = "rarely," 2 = "sometimes," and 3 = "often." In Magnus and Liu (2022), a random subset of 3000 complete responses was analyzed to illustrate the multidimensional hurdle graded response model (MH-GRM). Here we perform reliability calculations on the same data subset (i.e., $n = 3000$ and $m = 14$), which was retrieved from https://osf.io/frjm6/.

The MH-GRM was developed to account for an excessive amount of "never" endorsements in the data set (Magnus & Liu, 2022), and is a special case of the item response tree model (e.g., Jeon & De Boeck, 2016). Let each MV y_{ij} be the outcome of a two-stage decision process, in which $y_{ij} \in \{0, 1, 2, 3\}$, $j = 1, \ldots, 14$. The first stage represents the presence or absence of the symptom. Let $y_{ij}^{(1)} \in \{0, 1\}$ be a dichotomous indicator of the first stage such that $y_{ij}^{(1)} = 0$ maps onto $y_{ij} = 0$ (i.e., the "never" category), and $y_{ij}^{(1)} = 1$ maps onto $y_{ij} > 0$ (i.e., one of the three categories indicating symptom presence). The second stage determines the frequency of the symptom, which is encoded by a trichotomous indicator $y_{ij}^{(2)} \in \{1, 2, 3\}$. Assuming that the

symptom is present (i.e., $y_{ij}^{(1)} = 1$, set $y_{ij} = y_{ij}^{(2)}$). Conversely, $y_{ij}^{(1)}$ and $y_{ij}^{(2)}$ can be recoded from the item response y_{ij} by

$$y_{ij}^{(1)} = \begin{cases} 0, & y_{ij} = 0 \\ 1, & y_{ij} > 0 \end{cases} \quad \text{and} \quad y_{ij}^{(2)} = \begin{cases} \texttt{NA}, & y_{ij} = 0 \\ y_{ij}, & y_{ij} > 0 \end{cases} \tag{39}$$

If the symptom is absent (i.e., $y_{ij} = y_{ij}^{(1)} = 0$) no information about $y_{ij}^{(2)}$ can be inferred from y_{ij} and thus we code $y_{ij}^{(2)}$ as missing (i.e., `NA`) in Equation 39. The two-stage decision process is illustrated in Figure 2A.

Under the MH-GRM model, each individual i is characterized by two LVs, denoted by $\boldsymbol{\eta}_i = (\eta_{i1}, \eta_{i2})'$. The *susceptibility* LV η_{i1} is indicated by the symptom presence indicators $y_{i1}^{(1)}, \ldots, y_{i,14}^{(1)}$. The conditional distribution $\underline{y}_{ij}^{(1)} | \eta_{i1}$ is characterized by a 2PL model:

$$\mathbb{P}\{\underline{y}_{ij}^{(1)} = 1 | \eta_{i1}\} = \frac{\exp\left(\alpha_j^{(1)} + \beta_j^{(1)} \eta_{i1}\right)}{1 + \exp\left(\alpha_j^{(1)} + \beta_j^{(1)} \eta_{i1}\right)}, \tag{40}$$

in which $\alpha_j^{(1)}$ and $\beta_j^{(1)}$ represent the intercept and slope corresponding to the susceptibility LV η_{i1}. The *severity* LV η_{i2} is indicated only by the symptom frequency indicators $y_{i1}^{(2)}, \ldots, y_{i,14}^{(2)}$ through a graded response model:

$$\mathbb{P}\{\underline{y}_{ij}^{(2)} = k | \eta_{i2}\} = \frac{\exp\left(\alpha_{j,k-1}^{(2)} + \beta_j^{(2)} \eta_{i2}\right)}{1 + \exp\left(\alpha_{j,k-1}^{(2)} + \beta_j^{(2)} \eta_{i2}\right)} - \frac{\exp\left(\alpha_{jk}^{(2)} + \beta_j^{(2)} \eta_{i2}\right)}{1 + \exp\left(\alpha_{jk}^{(2)} + \beta_j^{(2)} \eta_{i2}\right)}, \; k = 1, 2, \text{ or } 3. \tag{41}$$

In Equation 41, $\alpha_{jk}^{(2)}$'s and $\beta_j^{(2)}$ represent the intercepts and slope corresponding to the severity LV η_{i2}, in which the intercepts must be in a descending order; i.e., $\infty = \alpha_{j0} > \alpha_{j1} > \alpha_{j2} > \alpha_{j3} = -\infty$. Additionally, η_{i1} and η_{i2} are allowed to covary (see Figure 2B). The estimated correlation between susceptibility and severity is 0.58, implying that they are related but distinctive constructs. Magnus and Liu (2022) also reported that EAP scores for susceptibility and severity exhibited different patterns in predicting health-related outcomes. For instance, when both sets of scores were used in a logistic regression to predict attempted suicide, only severity scores had a

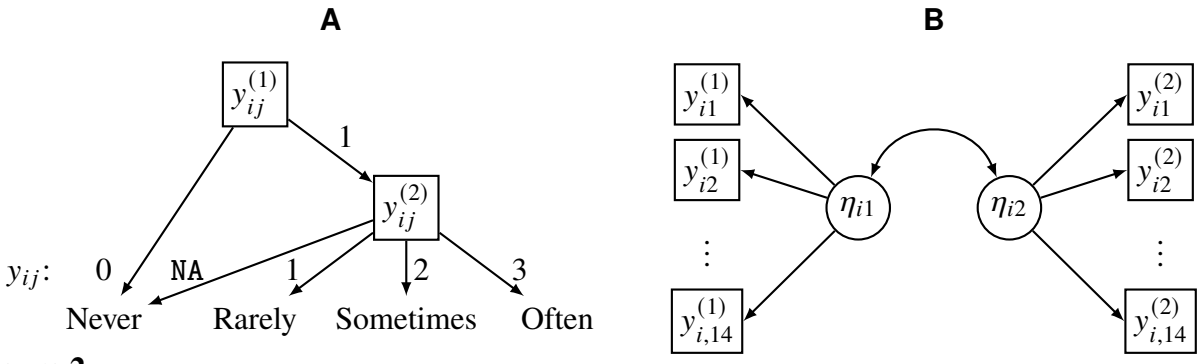

Figure 2

A: Response process. Each item response $y_{ij} \in \{0, 1, 2, 3\}$ is modeled by a two-stage decision process. The first stage represents presence of a symptom, mapping onto $y_{ij}^{(1)} \in \{0, 1\}$. The second stage represents symptom frequency, mapping onto $y_{ij}^{(2)} \in \{1, 2, 3\}$. $y_{ij} = 0$, 1, 2, and 3 an be rearranged to $(y_{ij}^{(1)}, y_{ij}^{(2)})' = (0, \texttt{NA})'$, $(1, 1)'$, $(1, 2)'$, and $(1, 3)'$, respectively, in which `NA` *denotes missing data. B: Path diagram. The susceptibility latent variable (LV) η_{i1} is indicated by the symptom frequency indicators $y_{i1}^{(1)}, \ldots, y_{i,14}^{(1)}$, and the severity LV η_{i2} is indicated by the symptom frequency indicators $y_{i1}^{(2)}, \ldots, y_{i,14}^{(2)}$. These two LVs are allowed to covary.*

significant partial effect.

To illustrate various reliability coefficients organized under the regression framework, we treat estimated parameters as known and ignore sampling variability because of the large sample size (n = 3000). Slightly different from Magnus and Liu (2022), in which model parameters are estimated by directly maximizing the marginal likelihood, we recoded the raw item response data using Equation 39 and fit the independent cluster model to the recoded data using the R package `mirt` (Chalmers, 2012). The software package implements an Expectation-Maximization algorithm for parameter estimation (Bock & Aitkin, 1981), and the default configuration of numerical quadrature and convergence criteria were used.

Measurement Decomposition

We first compute CTT reliability coefficients for three types of observed scores to illustrate measurement decompositions. Recall that under a measurement decomposition, we regress an observed score onto all LVs. The observed scores being considered here are: (a) the EAP score

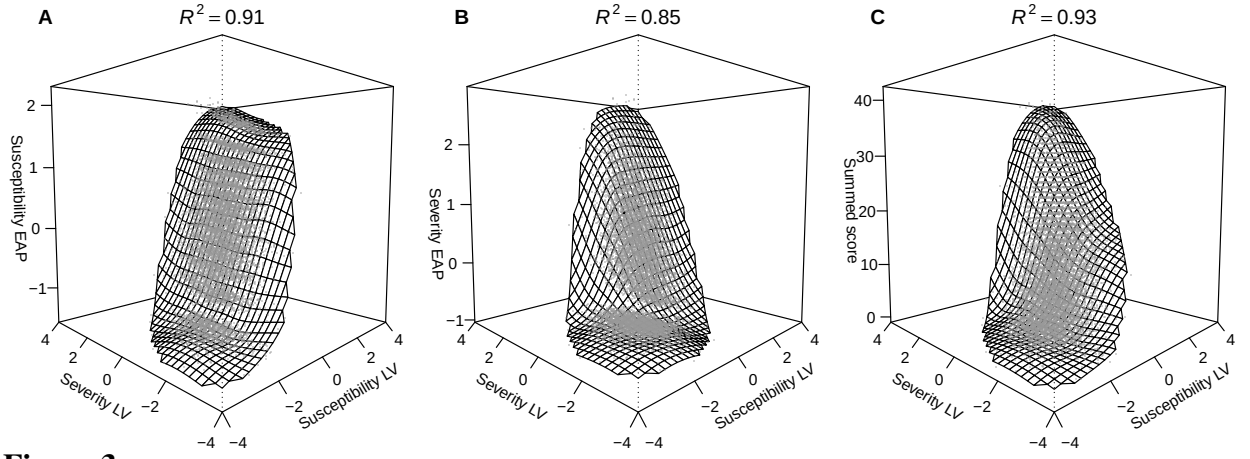

Figure 3

Illustration of measurement decomposition. The three panels correspond to the three types of observed scores: the expected a posteriori (EAP) score for susceptibility (A), the EAP score for severity (B), and the summed score (C). In each panel, the x-, y-, and z-axes represent susceptibility latent scores, the severity latent scores, and the observed score, respectively. A random subset of 10^4 *MC samples are selected and the corresponding scores are plotted (in gray dots). Predicted values of the nonparametric regression are presented by the mesh surface. The estimated* R^2 *(i.e., CTT reliability coefficient) is displayed on top of each plot.*

for the susceptibility factor $\mathbb{E}(\underline{\eta}_{i1}|\mathbf{y}_i)$, (b) the EAP score for the severity factor $\mathbb{E}(\underline{\eta}_{i2}|\mathbf{y}_i)$, and (c) the summed score $\mathbf{1}'_m\mathbf{y}_i$. The two LVs in the MH-GRM model are the susceptibility LV η_{i1} and the severity LV η_{i2}. The three nonparametric regressions obtained by Step 2 of our MC algorithm are visualized in separate panels of Figure 3. For a less dense visualization, we randomly selected a subset of 10^4 MC samples and generated three-dimensional scatter plots of the scores (i.e., the gray dots in Figure 3), in which the x- and y-axes represent the latent scores (i.e., the susceptibility and severity LVs) and the z-axis represents the observed score. The fitted values of the nonparametric regressions are depicted by mesh surfaces. The corresponding R^2 statistics, which are estimates of CTT reliability, are presented on top of the graphics.

The R^2 statistic obtained for the susceptibility EAP score is higher than that of the severity EAP score (i.e., 0.91 versus 0.85, respectively), indicating that the EAP score is more reliable than the CTT score. Moreover, the mesh surface in Figure 3A shows that the association between the susceptibility LV and its EAP score is strong, positive, and nonlinear, whereas the association

between severity LV and its EAP score is close to zero. In contrast, the severity EAP score is positively related to both susceptibility and severity except when both LVs have low values (see Figure 3B). Taken together, the susceptibility EAP score almost exclusively reflects the susceptibility LV. However, the severity EAP score reflects strong relations with the susceptibility and severity LVs. These observations can be gleaned from values of the estimated reliability coefficients and the wire frame plots made available via the MC algorithm.

The estimated CTT reliability for the summed score is 0.93, which is higher than the two EAP scores. This is expected because the fitted regression surface in Figure 3C suggests that both the susceptibility and severity LVs have strong partial effects on the summed score. Stated differently, the high reliability of the summed score is attributed to its strong, positive dependencies on both the susceptibility and severity LVs (relative to the two EAP scores). As a corollary of Theorem 4.4.3 in Lord and Novick (1968), CTT reliability of the summed score is bounded from below by coefficient alpha under the MH-GRM. In this example, the estimated coefficient alpha is 0.93, which is a very tight lower bound for our MC estimate of 0.93.

Prediction Decomposition

The two latent scores of interest are η_{i1} and η_{i2}, representing the constructs of susceptbility and severity, respectively. Recall that all the MVs in $\mathbf{y}_i$ serve as the observed scores in a prediction decomposition. Because the MVs are discrete, the prediction decomposition is a linear regression with 4^{14} coefficients in which MVs are dummy coded and we include lower order terms up to the 14-way interaction among the dummy variables. Because it is difficult to estimate and visualize this regression due to the large number of explanatory variables, we computed the response-pattern EAP scores using the `mirt` package and applied Step 2′ of the MC algorithm. We then plot the latent scores against these EAP scores which are composed of a random subset of 10^4 MC samples. In theory, the regression of a latent score on the associated EAP estimate should be linear with intercept zero and slope one (i.e., a unit-weight linear regression; see Equation 32).

The estimated PRMSE for the susceptibility and severity LVs are 0.88 and 0.72,

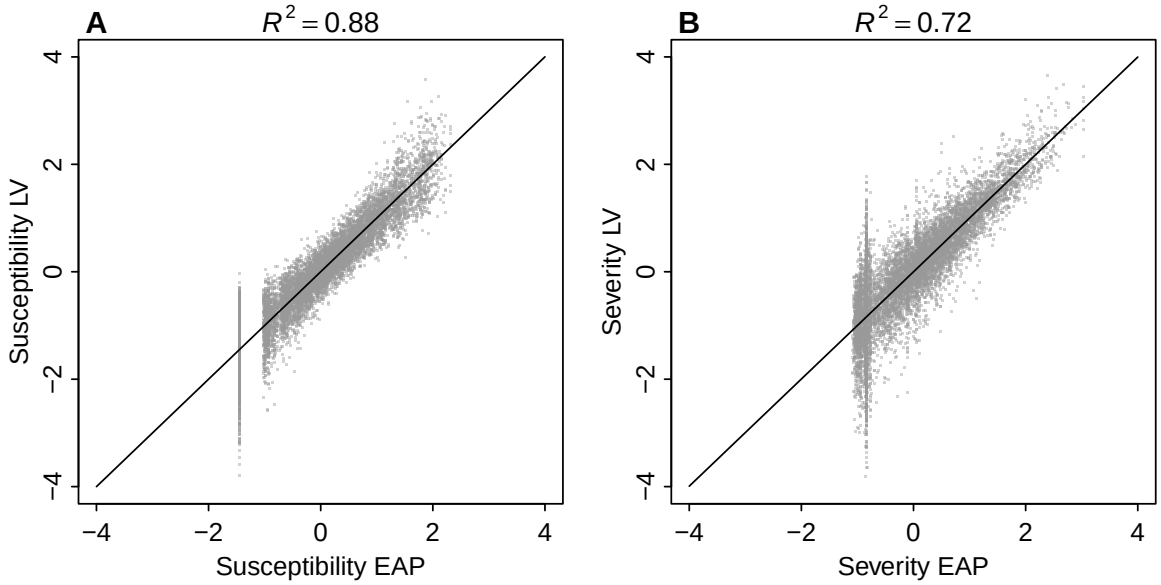

Figure 4

Illustration of prediction decomposition. The two panels correspond to the two latent scores: susceptibility latent scores η_{i1} (A) and the severity latent scores η_{i2} (B). In each panel, the y-axis represents the latent score, and the x-axis represents the associated expected a posteriori (EAP) observed score. A random subset of 10^4 *Monte Carlo (MC) samples are selected and the corresponding scores are plotted (in gray dots). Predicted values of the nonparametric regression are depicted as the solid curve. The estimated* R^2 *(i.e., PRMSE reliability coefficient) is displayed on top of each plot.*

respectively, implying that the MVs predict the construct of symptom presence better than the construct of symptom frequency under the MH-GRM. Observe also that the PRMSEs for the two LVs (0.88 for susceptibility and 0.72 for severity) are lower than their corresponding CTT reliabilities of the two EAP scores (0.91 for susceptibility and 0.85 for severity), which follows from Equation 31 of Kim (2012). The fitted linear regression has a zero intercept and a unit slope, presenting as diagonal straight lines (see Figure 4).

In sum, we presented an empirical example using a more complex measurement model for NCS-R depressive symptom data to illustrate approximate computations of CTT reliability and PRMSE with the proposed MC algorithm. Recall that these reliability coefficients can be obtained from regressing observed scores onto LVs (i.e., measurement decomposition) or regressing latent scores onto MVs (i.e., prediction decomposition). Estimated coefficients of determination

(namely, R^2s from the corresponding regressions) approximate CTT reliability and PRMSE. We emphasize that reliability is a property of the observed score under a measurement decomposition, but a property of the latent score under a prediction decomposition. Therefore, different reliability coefficients communicate distinct information and direct comparisons of reliability coefficients from different decompositions are discouraged.

Discussion

Consistent with the regression framework originated from McDonald (2011), reliability coefficients are coefficients of determination. Given a population measurement model that specifies the joint distribution of MVs and LVs, we can regress an observed score of interest onto all the LVs, leading to the measurement decomposition of the observed score (Equation 15). Alternatively, we can regress a latent score of interest onto all the MVs, leading to the prediction decomposition of the latent score (Equation 32). In a measurement decomposition, the predicted value of an observed score amounts to the underlying true score, and the corresponding coefficient of determination coincides with CTT reliability of the observed score (Lord & Novick, 1968). In a prediction decomposition, the predicted value of a latent score amounts to its EAP estimate, and the corresponding coefficient of determination coincides with PRMSE of the latent score (Haberman & Sinharay, 2010). Hence, we can alternatively view the measurement decomposition as a unit-weight linear regression of the observed score on its true score; similarly, the prediction decomposition can be viewed as a unit-weight linear regression of the latent score on its EAP estimate.

To compute approximate measures of reliability for computationally intractable situations, we introduced an MC procedure. With a large sample of simulated LVs and MVs, we can fit (nonparametric) regressions predicting either observed or latent scores and approximate the reliability coefficients by R^2 statistics. The MC algorithm allows one to obtain estimates for various types of reliability coefficients without directly evaluating integrals or large finite sums. For the linear one-factor and 2PL examples, we observed that MC-based reliability estimates are

nearly identical to their theoretical values. Furthermore, we demonstrated the adaptability of the proposed algorithm to more complex IRT models with an illustration using the NCS-R depressive symptom data. We also highlighted why different types of reliability coefficients (measurement versus prediction decomposition) should be expected to be different. Thus, when reporting reliability coefficients, it is essential to specify the underlying measurement model and the type of score (i.e., observed versus latent) used as an outcome in the regression framework.

We have re-introduced and expanded on McDonald's (2011) view on reliability with illustrations using a novel MC approach, leaving a number of related topics to be studied in the future. First, the performance of the MC procedure should be broadly examined under more extensive simulation settings (e.g., under different measurement models, sample sizes, and model error) as well as a broader set of empirical examples. It is anticipated that the optimal tuning of the MC algorithm (e.g., the number of MC samples M and the choice of nonparametric regression estimators) would differ from case to case. Second, in our empirical example, we treated estimated model parameters as population values, ignoring sampling variability. In practice, however, calibration studies of measurement models tend to have limited sample sizes. Thus, it would be important to quantify sampling variability inherent in estimated reliability coefficients based on our MC algorithm. Finally, to better understand the characteristics of various reliability coefficients, methodologists would need to develop benchmarks or recommendations on how these distinct measures of reliability might be qualitatively interpreted.

References

Alegria, M., Jackson, J. S., Kessler, R. C., & Takeuchi, D. (2001-2003). *Collaborative psychiatric epidemiology surveys (CPES), 2001-2003 [United States] (ICPSR 20240).* Inter-university Consortium for Political and Social Research [distributor], 2016-03-23. https://doi.org/10.3886/ICPSR20240.v8.

Birnbaum, A. (1968). Some latent train models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. doi: 10.1007/bf02293801

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence*, *30*(6), 505–514. doi: 10.1016/S0160-2896(02)00082-X

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Duxbury.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. doi: 10.18637/jss.v048.i06

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*, 300–315. doi: 10.1037/a0033805

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. doi: 10.1037/h0040957

Fox, J. (2015). *Applied regression analysis and generalized linear models*. SAGE Publications.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*, 209–227. doi: 10.1007/s11336-010-9158-4

Holland, P. W. (1990). On the sampling theory roundations of item response theory models. *Psychometrika*, *55*, 577–601. doi: 10.1007/bf02294609

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, *48*, 1070–1085. doi: 10.3758/s13428-015-0631-y

Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika*, *77*(1), 153–162. doi: /10.1007/s11336-011-9238-0

Liu, Y., & Pek, J. (2023). Summed versus estimated factor scores: Considering uncertainties when using observed scores. *Psychological Methods*. doi: 10.1037/met0000644

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*(2), 233–245. doi: 10.1007/bf02294018

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

Magnus, B. E., & Liu, Y. (2022). Symptom presence and symptom severity as unique indicators of psychopathology: An application of multidimensional zero-inflated and hurdle graded response models. *Educational and Psychological Measurement*, *82*(5), 938–966. doi: 10.1177/0013164421106182

McDonald, R. P. (1994). Testing for approximate dimensionality. In D. Levault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories of measurement: Problems and issues* (pp. 63–86). Ottawa, Canada: University of Ottawa.

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Lawrence Erlbaum.

McDonald, R. P. (2011). Measuring latent quantities. *Psychometrika*, *76*(4), 511–536. doi: 10.1007/s11336-011-9223-7

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395–1411. doi: 10.1037/pas0000754

Schmidt, F. L., & Hunter, J. E. (1999). *Theory testing and measurement error* (Vol. 27) (No. 3). Elsevier. doi: 10.1016/s0160-2896(99)00024-0

Stout, W. (2002). Psychometrics: From practice to theory and back: 15 years of nonparametric multidimensional IRT, DIF/test equity, and skills diagnostic assessment. *Psychometrika*, *67*, 485–518. doi: 10.1007/bf02295128

Thissen, D., & Steinberg, L. (2009). Item response theory. In R. Millsap & A. Maydeu-Olivares (Eds.), *The sage handbook of quantitative methods in psychology* (pp. 148–177). London: Sage Publications.

Thissen, D., & Thissen-Roe, A. (2022). Latent variable estimation in factor analysis and item response theory. *Chinese/English Journal of Educational Measurement and Evaluation*, *3*(3), Article 1. doi: 10.59863/optz4045

Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum.

Thomson, G. H. (1936). Some points of mathematical technique in the factorial analysis of ability. (27), 36–54. doi: 10.1037/h0062007

Thurstone, L. L. (1935). *The vectors on mind*. University of Chicago Press.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *65*(1), 95–114. doi: 10.1111/1467-9868.00374

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). CRC Press.