Educational and Psychological Measurement
72(1) 52–67

DOI: 10.1177/0013164411407315
http://epm.sagepub.com

Comparison of Reliability Measures Under Factor Analysis and Item Response Theory

Ying Cheng[1,2], Ke-Hai Yuan[1], and Cheng Liu[1]

Abstract

Reliability of test scores is one of the most pervasive psychometric concepts in measurement. Reliability coefficients based on a unifactor model for continuous indicators include maximal reliability ρ and an unweighted sum score–based ω, among many others. With increasing popularity of item response theory, a parallel reliability measure π has been introduced using the information function. This article studies the relationship among the three reliability coefficients. Exploiting the equivalency between item factor analysis and the normal ogive model, $\pi_{(2)}$ for dichotomous data is shown to be always smaller than ρ. Additional results imply that ω is typically greater than $\pi_{(2)}$ in practical conditions, though mathematically there is no dominant relationship between $\pi_{(2)}$ and ω. Further results indicate that, as the number of response categories increases, π can surpass ω. The reasons why π and ω fall short of ρ are also explored from an information gain/loss perspective. Implications of the findings on scale development and analysis are discussed.

Keywords

item response theory, factor analysis, information, reliability, maximal reliability

Introduction

Reliability is one of the most prevalent psychometric concepts in measurement. A search in PsycInfo of ''reliability'' yields 122,970 results. Meanwhile, it remains

[1]University of Notre Dame, Notre Dame, IN, USA
[2]Soochow University, Suzhou, Taiwan

Corresponding Author:
Ying Cheng, Department of Psychology, University of Notre Dame, 118 Haggar Hall, Notre Dame, IN 46556, USA
Email: ycheng4@nd.edu

one of the most concurrent topic in psychometric research. The search also returns more than 40 methodological articles on reliability since 2000. These articles address various aspects of reliability, including relations among different reliability indices (e.g., Zinbarg, Revelle, Yovel, & Li, 2005), the exact or asymptotic distribution of reliability indices (Kistner & Muller, 2004; van Zyl, Neudecker, & Nel, 2000; Yuan & Bentler, 2002), discussion on maximal reliability (Yuan & Bentler, 2002), and how reliability changes as a function of the number of item options (MacCann, 2004). Recently *Psychometrika* published a series of discussions on various reliability measures, their uses, common misconceptions about them, and theoretical issues remaining to be solved (Bentler, 2009; S. B. Green & Yang, 2009a, 2009b; Revelle & Zinbarg, 2009; Sijtsma, 2009a, 2009b).

Existing literature on estimation of reliability and related problems has focused predominantly on reliability of the composite scores of a set of items. Presumably these items are sources of information on the common, latent construct(s) (Penev & Raykov, 2006). Among these composites, the raw sum score has been used most extensively in empirical research. Cronbach's alpha is the most widely used measure of reliability of the raw sum score. But researchers have long been aware that alpha may not be the best choice. When the assumption of essential tau-equivalency is violated, alpha underestimates reliability at the population level. In fact, given that the unifactor model holds, alpha is a special case of a more broadly defined reliability measure, ω (McDonald, 1999), and $\alpha \leq \omega$. The equality holds if the items are essentially tau-equivalent.

Efforts on improving reliability through optimal weighted composite scores could be traced back to over half a century ago (Bentler, 1968; B. F. Green, 1952; Thompson, 1940). For recent discussions on the resulting ''maximal reliability'' ρ, please see Bentler (2007), Li (1997), Raykov (2004), and Raykov and Penev (2006). As suggested by its name, maximal reliability is the highest possible reliability that a test can achieve.[1] Under the same model, the maximal reliability ρ can also be achieved by finding the optimally weighted composite score. Details of ω and ρ will be provided later.

In recent years, with the increasing popularity of item response theory (IRT), reliability measures under IRT have caught much attention (Dimitrov, 2003; Kolen, Zeng, & Hanson, 1996; Mellenbergh, 1994, 1996; Nicewander, 1990, 1993; Samejima, 1994). For an up-to-date comprehensive summary, see Kim and Feldt (2010). In addition, researchers have been trying to relate the reliability measures defined under IRT to those under classical test theory (Andrich, 1982; Bechger, Maris, Verstralen, & Béguin, 2003; Nicewander, 1993). Meanwhile, the relationship between traditional reliability measures under classical test theory such as α, and those originated from factor analysis, such as ω and ρ, are well studied (e.g., Graham, 2006; Raykov, 1997). However, there is a lack of rigorous study on the relationship between reliability measures under IRT and factor analytic models in the literature. The purpose of this article is to establish the relationship between an IRT-based reliability coefficient (π) and two factor-analysis–based reliability measures (ω and ρ).

The relationship between ρ and ω has been well established. More specifically, ρ dominates ω and they are equal only when the ratio of factor loading and the unique variance holds constant across all items. Given that ρ is the maximal reliability, we may expect that π is also smaller than ρ. However, this has not been formally established in the literature. There also does not exist any study comparing π and ω. Actually, directly comparing π with ρ or ω is rather difficult. Our study on the relationship of π with ω and ρ will use the equivalency between the unidimensional normal ogive IRT models and the unifactor models. In particular, we will analytically show that, when the items are dichotomous, π is always smaller than ρ. We will explain the reasons on why π and ω fall short from an information loss/gain perspective.

Furthermore, using simulation and numerical examples, we will show that there is no dominant relationship between ω and π. Our results indicate that dichotomizing a continuous response usually results in very severe information loss—more severe than can be remedied by the use of factor score (as opposed to the simple sum score). Less extreme form of discretization such as adopting an over-2-point Likert-type scale can abate the information loss. The more response options there are, the more information we can retain, and consequently, the more reliable measure we can obtain about the underlying trait. The gain of information gradually decreases though, as the number of response categories increases.

Because estimates of reliability coefficients contain sampling errors, which typically vary as the data set changes, our study will be conducted at the population level. Thus, the obtained results hold for all samples from the same population and offer more valid information than results that hold at the sample level. Since sampling errors in any sensible estimates of the three reliability coefficients will go to zero as the sample size increases, we expect that the obtained relationships also hold in typical applications when sample sizes are not too small.

The following section gives a brief review of reliability under the unidimensional factor model and contains the details leading to ρ and ω. Then widely used IRT models and the reliability measure π defined under these models are introduced. The relationships between π and ρ, and π, ω, and ρ will be studied next. Implications to empirical research are discussed in the concluding section.

Reliability of a Composite Score

Without loss of generality, we consider the single-factor model:

$$X_j = \lambda_j \theta + e_j, \tag{1}$$

where X_j is a centered continuous standardized variable, θ the latent construct, and $e_j \sim \mathcal{N}(0,\psi_j^2)$. Thus, conditioning on θ and assuming unit variance for θ, we have $X_j|\theta \sim \mathcal{N}(\lambda_j\theta,\psi_j^2)$, where $\psi_j^2 = 1 - \lambda_j^2$. The model is a special case of the item factor analysis model (e.g., Bock & Aitkin, 1981; Muraki & Engelhard, 1985), where the items are congeneric. Suppose we have m items and let

$$X = \sum_{j=1}^{m} X_j = T + e = \sum_{j=1}^{m} \lambda_j \theta + \sum_{j=1}^{m} e_j, \tag{2}$$

where $T = \sum_{j=1}^{m} \lambda_j \theta$ is known as the true score, and $e = \sum_{j=1}^{m} e_j$ is the error term. By the definition of reliability, given the model defined in Equation (1),

$$\omega = \frac{(\sum_{j=1}^{m} \lambda_j)^2}{[(\sum_{j=1}^{m} \lambda_j)^2 + \sum_{j=1}^{m} \psi_j^2]} = \frac{1}{1 + [(\sum_{j=1}^{m} \lambda_j)^2 / \sum_{j=1}^{m} \psi_j^2]^{-1}}, \tag{3}$$

which is the ratio of the variance of the true score over the variance of the unweighted composite score X. Under the model defined in Equation (1), the widely used Cronbach's alpha is a lower bound of ω. It is equal to ω only when the items load equally on the underlying factor, that is, $\lambda_j = \lambda$ for $j = 1, 2, \ldots, m$ (McDonald, 1999).

It is known that the X in Equation (2) is not the best estimate of the true score. Intuitively, any type of aggregation of item scores may result in loss of information. The unweighted composite score X neglects the fact that people with the same sum score can have completely different response patterns across items. In fact, for the same test, the maximum likehood (ML) factor score, $\hat{\theta}_{ML}$, enjoys the maximal reliability (see also Bentler, 1968; Li, 1997; McDonald, 1999; Raykov, 2004):

$$\rho = \frac{1}{1 + [\sum_{j=1}^{m} (\lambda_j^2 / \psi_j^2)]^{-1}}. \tag{4}$$

It is well known that ρ is at least as large as ω, that is, $\rho \geq \omega$.

Item Response Theory Models

So far, our discussion is limited to the case where the item responses are continuous. In behavioral research, however, the responses are often discrete. For instance, in survey research, likely response options are ''agree/disagree,'' or Likert-scale type of response options with multiple categories, such as ''strongly disagree, disagree, neutral, agree, strongly agree.'' The former represents a binary response, whereas the latter features graded response. Both cases can be considered as some form of discretization of a continuous response.

The IRT models directly relate the discrete responses to an underlying latent factor. For instance, the two-parameter normal ogive model specifies the relation of a dichotomized response Y_j and the latent factor θ as

$$P(Y_j = 1|\theta) = \int_{-\infty}^{\alpha_j(\theta - \beta_j)} \phi(t)dt, \tag{5}$$

where $\phi(t)$ is the density of $\mathcal{N}(0,1)$, and α_j and β_j are referred to as the discrimination and difficulty parameters of item j, respectively. These terms originate from the achievement-testing context, where $P_j(\theta) = P(Y_j = 1|\theta)$ represents the probability of a correct response. The ''discrimination'' parameter then characterizes how powerful an item is in telling apart high- versus low-achieving students, and the ''difficulty'' parameter characterizes how hard an item is. We will use $Q_j(\theta) = 1 - P_j(\theta)$ to denote the probability of an incorrect response or nonendorsement. Because of the computational complexity of the normal ogive model in Equation (5), the two-parameter logistic model (Birnbaum, 1968):

$$P_j(\theta) = \frac{\exp[D\alpha_j(\theta - \beta_j)]}{1 + \exp[D\alpha_j(\theta - \beta_j)]} \tag{6}$$

is frequently used in applied research, where $D = 1.701$ makes the logistic function and the normal ogive differ by less than .01 over the entire range of θ. Thus, models shown in Eqquations (5) and (6) can be treated as indistinguishable.

Takane and de Leeuw (1987) showed that, when θ is normally distributed, the normal ogive IRT models are equivalent to the item factor analysis model.[2] The equivalence can be established between the two-parameter normal ogive model and the item factor analysis model by setting

$$\alpha_j = \frac{\lambda_j}{\psi_j}, \tag{7}$$

and

$$\beta_j = \frac{h_j}{\psi_j D\alpha_j} = \frac{h_j}{D\lambda_j},$$

where h_j is the dichotomizing threshold for a continuous response X_j: When $X_j > h_j$, $Y_j = 1$; otherwise $Y_j = 0$.

For responses of ordered categories, Samejima (1969) proposed the graded response model (GRM). Suppose the continuous response to item j is discretized by g thresholds, which results in $g + 1$ response categories. Assume the score categories are 0, 1, 2, . . ., g. The normal ogive GRM characterizes the probability of $Y_j = y$ as

$$P_{jy}(\theta) = P^*_{jy}(\theta) - P^*_{j(y+1)}(\theta), \tag{8}$$

for $y = 0, 1, 2, \ldots g$, where

$$P^*_{jy}(\theta) = \int_{-\infty}^{\alpha_j(\theta - \beta_{jy})} \phi(t)dt \tag{9}$$

defines the probability of endorsing response category y or higher. The β_{jy}s are category location or step parameters, which can be viewed as the boundary of two adjacent

response categories. It is also defined that $P^*_{j0}(\theta) = 1$ and $P^*_{j(g+1)}(\theta) = 0$. Similarly, the normal ogive version of the GRM is equivalent to the factor analysis of ordered categorical data. The equivalence can be established by having Equation (7) and

$$\beta_{jy} = \frac{h_{jy}}{D\lambda_j},$$

for $y = 0, 1, 2, \ldots, g$, where h_{jy} are the discretizing thresholds: When X_j falls between h_{jy} and $h_{j,y+1}$, $Y_j = y$.

Because the logistic versions of the IRT models are good approximation to their normal ogive counterparts, the equivalence also holds approximately between the two-parameter logistic and the logistic GRM models and their item factor analysis counterparts, a property that will be exploited in our study of the three reliability coefficients in the following sections.

Information and Reliability

In this section, we will connect the three reliability coefficients ρ, ω, and π using the concept of information. Information is closely related to maximum likelihood estimates (MLEs) in statistical inference. With a single parameter, it is defined as the negative expected value of the second derivative of the log likelihood function. The variance of the MLE is (approximately) given by the inverse of the information. In this section, we use information in a broader sense by equating it with the inverse of a variance even when the parameter estimate is not an MLE.

Let θ and $\hat{\theta}_{\mathrm{ML}}$ denote the latent trait and its MLE, respectively. Then $\hat{\theta}_{\mathrm{ML}} = \theta + \varepsilon$, where ε is the error term whose variance is given by the inverse of information, I. Note that the variance of θ is 1. It then follows from the definition that the reliability of $\hat{\theta}_{\mathrm{ML}}$ is

$$\text{Reliability} = \frac{I}{1+I}, \tag{10}$$

Following from Equation (2), let

$$\hat{\theta}_{\mathrm{UW}} = \frac{X}{\sum_{j=1}^{m} \lambda_j} = \theta + \frac{\sum_{j=1}^{m} e_j}{\sum_{j=1}^{m} \lambda_j}$$

be the factor score estimate based on the unweighted composite score X, where the denominator $\sum_{j=1}^{m} \lambda_j$ is applied to make the variance of the true score equal to 1.0. Then the information associated with $\hat{\theta}_{\mathrm{UW}}$ can be expressed as (McDonald, 1999)

$$I_{\mathrm{UW}}(X) = \frac{(\sum_{j=1}^{m} \lambda_j)^2}{\sum_{j=1}^{m} \psi_j^2}. \tag{11}$$

Let $w_j = \lambda_j / [\psi_j^2 \sum_{j=1}^{m} (\lambda_j^2 / \psi_j^2)]$. Then $\hat{\theta}_{\mathrm{ML}} = \sum_{j=1}^{m} w_j X_j$ is the MLE of θ based on the model (Equation 1), whose information is

$$I_{\mathrm{W}}(\hat{\theta}_{\mathrm{ML}}) = \sum_{j=1}^{m} \frac{\lambda_j^2}{\psi_j^2}, \tag{12}$$

where the subscript W means ''weighted'' in contrast to the unweighted simple summation. Notice that $\hat{\theta}_{\mathrm{ML}}$ is just the Bartlett factor score (Bartholomew, Deary, & Lawn, 2009).

With the information just identified, reliability for each of the estimators can be obtained by Equation (10). In particular, Equation (3) is obtained when plugging Equation (11) into Equation (10) and Equation (4) is obtained when plugging Equation (12) into Eq. 10. Thus, the reduction in reliability from ρ to ω is solely caused by the loss of information when the raw sum score is used. When $\hat{\theta}_{\mathrm{ML}}$ is used, response pattern is taken into account. When items load differentially on the underlying factor, two response patterns with the same raw sum score could lead to two distinct factor scores. Therefore, using $\hat{\theta}_{\mathrm{ML}}$ instead of X, more information can be retained. Similarly, we can also obtain the reliability π for IRT models using Equation (10). To do that, we need to obtain the corresponding information first.

Unlike for the linear model in Equation (1), information in IRT is a conditional concept. In other words, it is a function of θ. Specifically, when the continuous response is dichotomized as shown in the previous section, the information of each item contained in the factor score $\hat{\theta}_{\mathrm{ML}}$ based on the two-parameter logistic model is

$$I_{j,\theta}(\hat{\theta}_{\mathrm{ML}}) = D^2 \alpha_j^2 P_j(\theta) Q_j(\theta), \tag{13}$$

where $P_j(\theta)$ is defined in Equation (6), and $Q_j(\theta) = 1 - P_j(\theta)$. Under the condition of local independence, that is, when item responses are independent conditioning on θ, test information is a straight sum of item information,

$$I_{(2,\theta)}(\hat{\theta}_{\mathrm{ML}}) = \sum_{j=1}^{m} I_{j,\theta}(\hat{\theta}_{\mathrm{ML}}), \tag{14}$$

where the 2 in the subscript refers to the fact that this test information function is derived from the dichotomous IRT model. For the rest of the article, we will use $I_{(2,\theta)}$ as a shorthand for $I_{(2,\theta)}(\hat{\theta}_{\mathrm{ML}})$.

Equations (14) and (12) differ in two aspects: First, Equation (12) is based on continuous responses, whereas Equation (14) is based on dichotomous responses. Second,

Equation (12) is not a function of the latent trait, whereas Equation (14) apparently does depend on θ. This suggests that the IRT-based reliability measure needs to be ''globalized.'' In other words, θ needs to be integrated throughout the whole range of θ to be comparable to those based on the factor analysis model (Nicewander, 1993). Define the reliability under dichotomous IRT model as $\pi_{(2)}$. The appendix contains the details leading to

$$\pi_{(2)} = \frac{1}{1 + E_\theta[1/I_{(2,\theta)}]}. \tag{15}$$

This global reliability coefficient has also been discussed in Samejima (1994). Note the functional equivalence between Equations (10) and (15) when we rewrite

$$\text{Reliability} = \frac{1}{1 + 1/I}.$$

For the relationship between $\pi_{(2)}$ and ρ, we have the result

$$\rho > \pi_{(2)}. \tag{16}$$

Because the details leading to Equation (16) are quite complicated, we put them in the appendix. The result in Equation (16) implies that, even though $\hat{\theta}_{\text{ML}}$ is used, dichotomization leads to information loss in general. Notice that dichotomization is the most severe form of discretization. We would expect that π will be closer to ρ when adopting a Likert-type scale or graded responses with less information loss (Samejima, 1969). We will further evaluate the change of π as the number of response category increases in the next section. To do that, we need to have the information function corresponding to the GRM defined in Equations (8) and (9). Parallel to Equation (13), the information from the *j*th item is given by

$$I_{j,\theta}(\hat{\theta}_{\text{ML}}) = \sum_{y=0}^{g} \frac{[P'_{jy}(\theta)]^2}{P_{jy}(\theta)} = \sum_{y=1}^{g} \frac{\left\{[P^*_{jy}(\theta)]' - [P^*_{j(y+1)}(\theta)]'\right\}^2}{[P^*_{jy}(\theta) - P^*_{j(y+1)}(\theta)]}.$$

Similarly, the test information is simply the sum of the individual item information values. The test information is denoted by $I_{(g+1,\theta)}$ for a test based on the GRM with *g* discretizing thresholds. Following the same logic as in the case of dichotomous responses, we have

$$\pi_{(g+1)} = \frac{1}{1 + E_\theta[1/I_{(g+1,\theta)}]}.$$

When $g \to \infty$, the response is essentially continuous, and there is no information loss because of discretization. Then we would expect that $\pi_{(g+1)} \to \rho$ when $g \to \infty$. This will be examined in our study in the next section.

In summary, $\pi_{(2)}$ and ω are both smaller than ρ because of information loss. For $\pi_{(2)}$, it is because of dichotomization of continuous item responses; for ω, it is because of ignoring response pattern.

Comparison of ω and π

We have mathematically established that $\rho > \pi$ when items are dichotomous in the previous section. It is also well known that $\rho \geq \omega$. In this section, we study the relationship of ω and π using simulation and analysis. Since our study is at the population level, the simulation is for randomly chosen parameter values, not randomly generated data or samples.

Figure 1 contains ρ, ω, and $\pi_{(2)}$ from 1,000 runs of a 30-item test, where the parameters α_j are generated from the uniform distribution on the interval (0.5, 1.5), and β_j are generated from the uniform distribution on the interval (−3.0, 3.0). The ranges of α and β are chosen to reflect those of typical psychological and educational assessments. The parameters λ_j and $\psi_j = (1 - \lambda_j^2)^{1/2}$ are obtained according to the parameter equivalency shown in the ''Item Response Theory Models'' section. In the left panel, the blue points represent ρ, red points ω, and green points $\pi_{(2)}$. Clearly in that plot $\rho > \omega > \pi_{(2)}$. The other three panels, from left to right, represent $\rho - \omega$, $\rho - \pi_{(2)}$, and $\omega - \pi_{(2)}$, respectively. Again these three panels show positive values, which suggests that $\rho > \omega > \pi_{(2)}$. Since the simulation covers a wide range of α and β, and the ranges of these parameters mirror what are usually seen with typical psychological and educational assessments, it suggests that in most cases dichotomizing continuous responses results in more severe information loss than aggregating item-level responses.

However, ω is not necessarily greater than $\pi_{(2)}$ under all conditions. Consider a 30-item test in which all the items have difficulty parameter $\beta = 0$. Moreover, 29 of these items have $\alpha = 0.05$, whereas the last one's $\alpha = 1.5$. Such a test yields $\omega = 0.15$, and $\pi_{(2)} = 0.29$. This counterexample shows that the relative size of ω and $\pi_{(2)}$ depends on the item parameters. In reality, tests with parameters as in the counterexample are rare because it contains mostly very bad items. But it demonstrates that mathematically $\pi_{(2)}$ can be larger than ω, even by a fairly large margin. This happens because when the α values vary a lot among items, ω will get hard hit failing to give more weights to the items with higher α values. In that case, $\pi_{(2)}$ could be larger than ω.

As mentioned before, dichotomization is the most severe form of discretization. When the number of thresholds g increases, the reliability $\pi_{(g+1)}$ increases. Figure 2 shows how $\pi_{(g+1)}$ increases with g. Again, it is based on a 30-item test whose α parameters follow the uniform distribution on the interval (0.5, 1.5). All the step/category location parameters are generated from the uniform distribution on the interval (−3.0, 3.0). The parameters λ_j and $\psi_j = (1 - \lambda_j^2)^{1/2}$ are again obtained according to the parameter equivalency shown in the ''Item Response Theory Models'' section. Ten replications are run to show the effect of locations of the β_gs. Clearly, as g rises, $\pi_{(g+1)}$ can surpass ω, and finally converges to ρ as $g \to \infty$. This happens because when $g \to \infty$, we are essentially getting the continuous response back. Meanwhile,

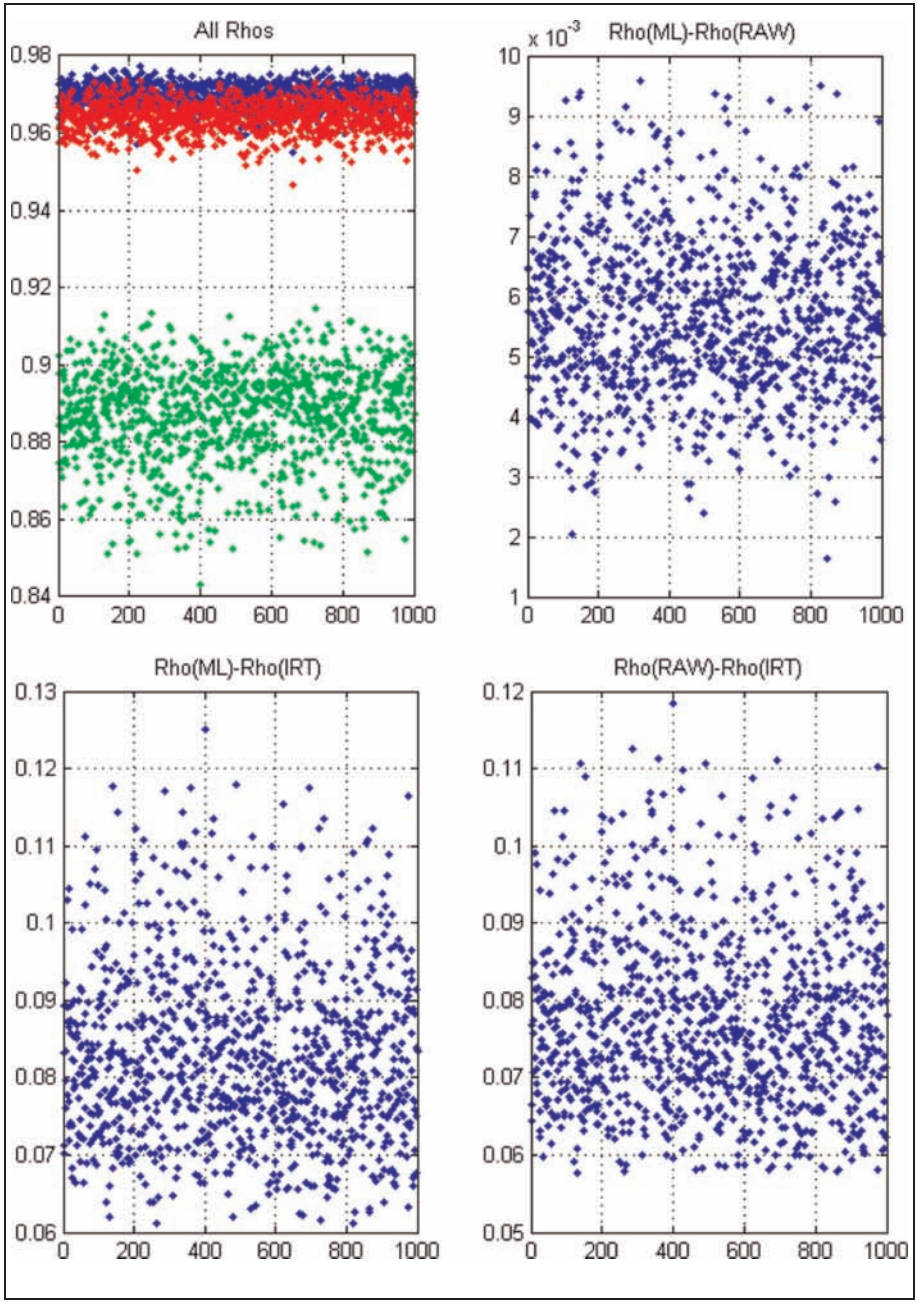

Figure 1. Comparison of ρ, ω, and $\pi_{(2)}$
Note. ML = maximum likelihood; IRT = item response theory.

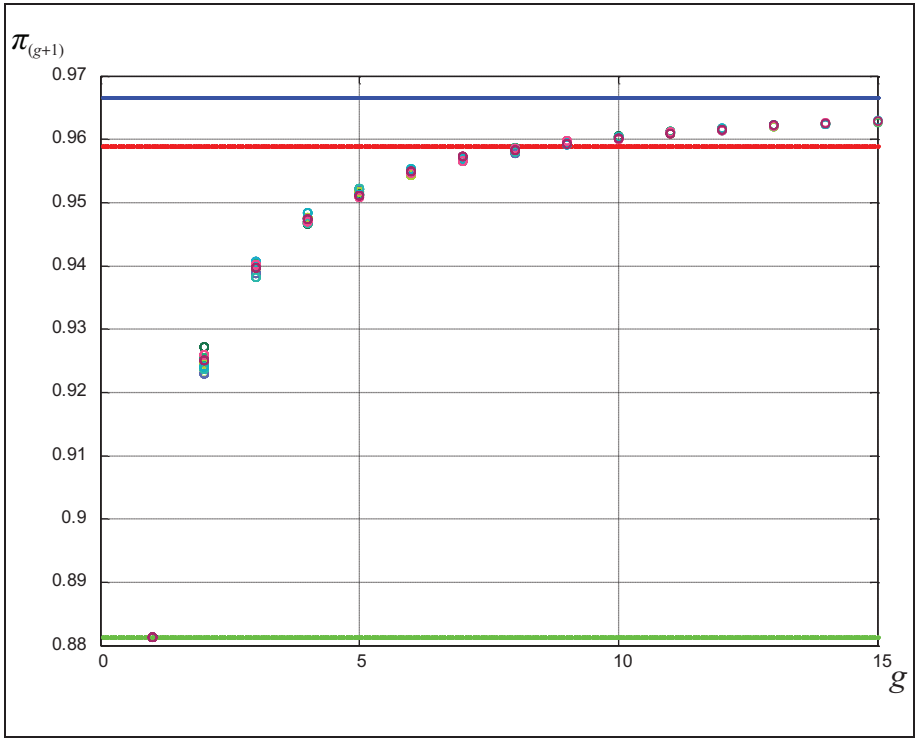

Figure 2. The relationship between $\pi_{(g+1)}$ and the number of thresholds g when the loadings are homogeneous

the locations of the β_gs affect the reliability but not to the same extent as the number of g. In this case, it needs a g of at least 10, implying at least 11 response categories, for $\pi_{(g+1)}$ to catch up with ω.

Figure 3 replicates what is shown in Figure 2, except that we are generating α from the uniform distribution on the interval (0, 2.5), instead of on the interval (0.5, 1.5). The general trend is the same, but now it only needs a $g \geq 5$ for $\pi_{(g+1)}$ to catch up with ω. The reason, again, is because Figure 2 is based on a test with items of more homogeneous loadings. In that case, ω is close to ρ, which is the maximal reliability. Our results indicate that when the loadings are more heterogeneous, ω suffers more, and it is easier for $\pi_{(g+1)}$ to catch up with ω.

Conclusion and Discussion

In this article, we studied the relationship among three reliability measures ρ, ω, and π. The first two are derived under factor analytic models, whereas the last, π, is defined on the basis of the IRT models. A special case of π is $\pi_{(2)}$ for binary responses, which

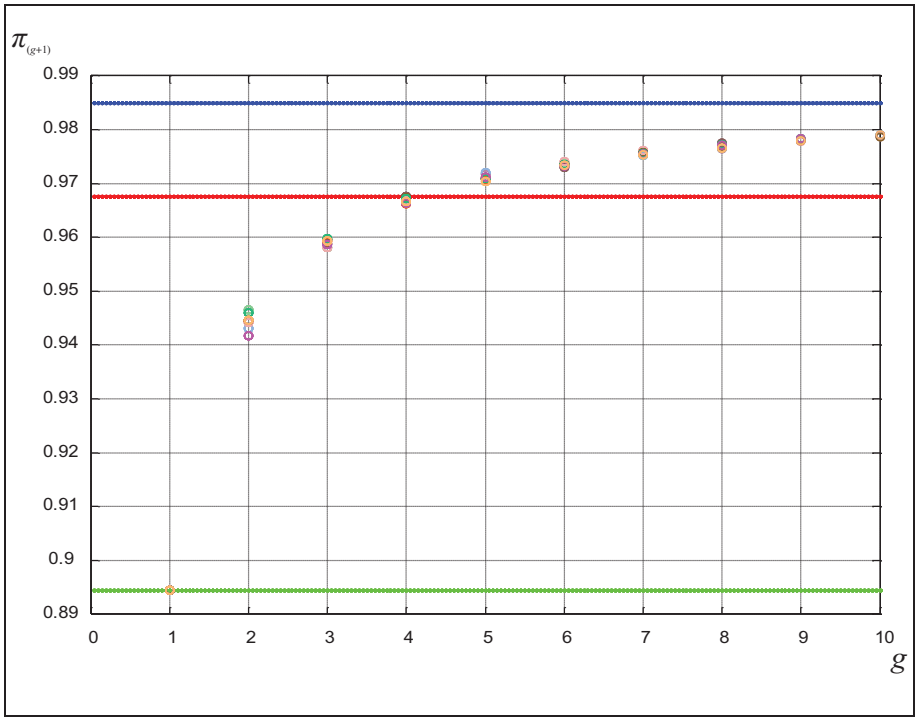

Figure 3. The relationship between $\pi_{(g+1)}$ and the number of thresholds g when the loadings are heterogeneous

is most widely used in cognitive testing. We showed algebraically that ρ is greater than $\pi_{(2)}$. We also showed that, when item parameters are within typical ranges of psychological and educational assessments, $\pi_{(2)}$ is smaller than ω. We also provided a numerical example indicating that there is no dominant relationship between $\pi_{(2)}$ and ω. Using simulation, we further demonstrated that, as the number of response options increases, π can exceed ω in practice.

We also explored the reasons why $\pi_{(2)}$ and ω fall short of ρ from an information gain/loss perspective. On $\pi_{(2)}$'s part, the information loss comes from dichotomizing continuous responses; on ω's part, from using unweighted sum score, which ignores the fact that response patterns may still differ given the same sum score. Less extreme form of discretization such as adopting an over-2-point Likert-type scale can abate the information loss in π; whereas using the MLE of θ in the linear model (Equation 1), which is just the Bartlett's factor score, will abate the loss in ω.

Multiple-choice items are widely used in psychological and educational assessments. There are many incidents where the responses are forced to be binary (e.g., agree vs. disagree, yes vs. no, or correct vs. incorrect). The advantage is that scoring becomes easy when the responses are binary, especially with Scantrons, but by doing

so we lose a substantial amount of information. Using an ML estimate of the underlying factor helps keep information, but the gain most of the time is still not big enough to abate the loss due to dichotomization.

On the other hand, when the items have very different factor loadings, that is, when the item discrimination parameters vary considerably, test reliability suffers from using an unweighted sum score. In summary, the best solution is to keep as many response categories as possible and to use the ML factor score. The benefit of adding more response categories, however, will gradually vanish when g increases. In other words, reliability does not respond to the number of categories linearly. After having a certain number of response options, it may not be worth adding more. When items load to a similar extent on the latent factor, it takes more response categories for π to catch up with ω. If the loadings are heterogeneous, it takes fewer response categories for π to outreach ω.

Nonetheless, our study is limited in several aspects. First, we only studied graded response models which assume that the item responses are ordered. There are other types of polytomous IRT models, such as the nominal model (Bock, 1972), which do not assume that the item responses are ordinal. The nominal model is widely used in attitude instrument or survey. Second, only unidimensional models are studied. We expect the same pattern of relationship will hold among the reliability coefficients with multidimensional models, when appropriate adjustments are made. For instance, ω would not be a good measure of reliability, when a multidimensional IRT model fits the data. Instead, more general reliability measures based on multiple factors as defined in Bentler (2007) or Revelle and Zinbarg (2009) would be appropriate references of comparison.

Appendix

This appendix contains the details leading to the results in Equations (15) and (16). It is known that asymptotically $\hat{\theta}_{ML} \sim \mathcal{N}(\theta,\sigma^2(\theta))$ with $E(\hat{\theta}_{ML}|\theta) = \theta$, and given a dichotomous IRT model,

$$\mathrm{Var}(\hat{\theta}_{ML}|\theta) = \frac{1}{I_{(2,\theta)}}.$$

It follows from the variance decomposition formula

$$\mathrm{Var}(\hat{\theta}_{ML}) = E[\mathrm{Var}(\hat{\theta}_{ML}|\theta)] + \mathrm{Var}[E(\hat{\theta}_{ML}|\theta)]$$

that

$$\mathrm{Var}(\hat{\theta}_{ML}) = E[1/I_{(2,\theta)}] + \mathrm{Var}(\theta). \tag{A1}$$

Equation (15) follows from Equation (A1) by noticing that $\mathrm{Var}(\theta) = 1$. Substituting $I_{(2,\theta)}$ in Equation (A1) by Equations (13) and (14) yields

$$\pi_{(2)} = \frac{1}{1 + E[1/\sum_{j=1}^{m} D^2\alpha_j^2 P_j(\theta)Q_j(\theta))]}.$$

Since $P_j(\theta)Q_j(\theta) \leq 1/4$,

$$\pi_{(2)} \leq \frac{1}{1 + 4/\sum_{j=1}^{m} D^2\alpha_j^2}.$$

Notice that $D \approx 1.701$,

$$\pi_{(2)} < \frac{1}{1 + 1/\sum_{j=1}^{m} \alpha_j^2}. \quad \text{(A2)}$$

It follows from Equations (4) and (7) that $\rho = 1/[1 + (\sum_{j=1}^{m} \alpha_j^2)^{-1}]$. Thus, Equation (A2) is just the result in Equation (16).

Acknowledgments

Part of the work was carried out while the first author was a summer visiting scholar at Soochow University.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Ke-Hai Yuan's work on this project is supported by Grant No. 4R44CA137841 from the National Cancer Institute.

Notes

1. Theoretically, using nonlinear combinations/functions of the ''components'' can lead to even higher reliability than the maximal reliability defined in this article. For details, see Knott and Bartholomew (1993). But the nonlinear composite score is beyond the scope of this article.
2. For the special case when θ is unidimensional and the item responses are binary, the equivalency was established by Lord and Novick (1968).

References

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index and the Guttman scale response pattern. *Education Research and Perspectives*, *9*, 95-104.

Bartholomew, D. J., Deary, I. J., & Lawn, M. (2009). The origin of factor scores: Spearman, Thomson and Bartlett. *British Journal of Mathematical and Statistical Psychology*, *62*, 569-582.

Bechger T. M., Maris, G., Verstralen, H., & Béguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, *27*, 319-334.

Bentler, P. M. (1968). Alpha-maximized factor analysis (Alphamax): Its relation to alpha and canonical factor analysis. *Psychometrika*, *33*, 335-345.

Bentler, P. M. (2007). Covariance structure models for maximal reliability of unit-weighted composites. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 1-19). Amsterdam, Netherlands: Elsevier North-Holland.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*, 137-143.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee s ability. In F.M Lord and M. R. Novick (eds.), *Statistical theories of mental test scores*, (pp. 397-472), Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443-458.

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

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, *66*, 930-944.

Green, B. F. (1952). A note on the calculation of weights for maximum battery reliability. *Psychometrika*, *17*, 57-61.

Green, S. B., & Yang, Y. (2009a). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*, 121-135.

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

Kim, S., & Feldt, L. S. (2010). The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Education Review*, *11*, 179-188.

Kistner, E. O., & Muller, K. E. (2004). Exact distributions of intraclass correlation and Cronbach's alpha with Gaussian data and general covariance. *Psychometrika*, *69*, 459-474.

Knott, M., & Bartholomew, D. J. (1993). Constructing measures with maximum reliability. *Psychometrika*, *58*, 331-338.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, *33*, 129-140.

Li, H. (1997). A unifying expression for the maximal reliability of a linear composite. *Psychometrika*, *62*, 245-249.

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

MacCann, R. G. (2004). Reliability as a function of the number of item options derived from the ''knowledge or random guessing'' model. *Psychometrika*, *69*, 147-157.

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

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*, 223-236.

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

Muraki, E., & Engelhard, G. (1985). Full-information item factor analysis: Applications of EAP scors. *Applied Psychological Measurement*, *9*, 417-430.

Nicewander, W. A. (1990). A latent-trait based reliability estimate and upper bound. *Psychometrika*, *55*, 65-74.

Nicewander, W. A. (1993). Some relationships between the information function of IRT and the signal/noise ratio and reliability coefficient of classical test theory. *Psychometrika*, *58*, 139-141.

Penev, S., & Raykov, T. (2006). Maximal reliability and power in covariance structure models. *British Journal of Mathematical and Statistical Psychology*, *59*, 75-87.

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, *32*, 329-353.

Raykov, T. (2004). Estimation of maximal reliability: A note on a covariance structure modeling approach. *British Journal of Mathematical and Statistical Psychology*, *57*, 21-27.

Raykov, T., & Penev, S. (2006). A direct method for obtaining approximate standard error and confidence interval for maximal reliability for composites with congeneric measures. *Multivariate Behavioral Research*, *41*, 15-28.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*, 145-154.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society.

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

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

Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika*, *74*, 169-173.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393-408.

Thompson, G. H. (1940). Weighting for battery reliability and prediction. *British Journal of Mathematical and Statistical Psychology*, *30*, 357-360.

van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika*, *65*, 271-280.

Yuan, K.-H., & Bentler, P. M. (2002). On robustness of the normal-theory based asymptotic distributions of three reliability coefficient estimates. *Psychometrika*, *67*, 251-259.

Zinbarg, R., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_h: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 122-133.