Conceptions of reliability revisited and practical recommendations

Sijtsma, K.; van der Ark, L.A.

DOI
10.1097/NNR.0000000000000077
Publication date
2015
Document Version
Final published version
Published in
Nursing research

Conceptions of Reliability Revisited and Practical Recommendations

Klaas Sijtsma ▼ L. Andries van der Ark

We discuss reliability definitions from the perspectives of classical test theory, factor analysis, and generalizability theory. For each method, we discuss the rationale, the estimation of reliability, and the goodness of fit of the model that defines the reliability coefficient to the data. Similarities and differences in the three approaches are highlighted. Finally, we provide a computational example using generated data to illustrate the differences among the different reliability methods.

Key Words: classical test theory reliability • factor analysis reliability • generalizability coefficients • test score reliability

Nursing Research, March/April 2015, Vol 64, No 2, 128–136

This article discusses different reliability conceptions for measurement of attributes in the health and nursing sciences. Example attributes are the pain patients experience who suffer from burn wounds (de Jong, Bremer, Schouten, Tuinebreijer, & Faber, 2005); aspects of health-related quality of life such as physical functioning, general health perceptions, vitality, and social functioning (Gandek, Sinclair, Kosinski, & Ware, 2004); adherence to medication and lifestyle for patients experiencing hypertension (Ma, Chen, You, Luo, & Xing, 2012); and nursing skills (National Council of State Boards of Nursing, 2009). Like investigators in fields such as psychology, sociology, political science, and marketing, researchers in the health and nursing sciences use multi-item measurement instruments to measure typical attributes. Items can be printed rating-scale statements (questionnaire), oral questions (interview), or structured observations (Sijtsma, in press).

Reliability and validity are the two basic properties of measurement values used in the health and nursing sciences. Measurement instruments produce measurement values or test scores that represent the measured attribute. Reliability refers to the degree to which a set of measurement values can be repeated under precisely the same measurement conditions, thus reflecting the fundamental question in statistics: "What would happen with the results if I could do the research over again?" Validity is the degree to which the measurement reflects the intended attribute, but it also refers to the suitability of the measurement value for a particular use, such as the identification of patients experiencing pain because of burn wounds for a particular therapy or the assessment of the necessity that people showing low levels of social functioning participate in social neighborhood programs. This article focuses on reliability. For views on validity and discussions, the reader may consult Wainer and Braun (1988), Messick (1989), Lissitz (2009), Kane (2013), and Markus and Borsboom (2013).

We discuss three psychometric approaches to reliability: classical test theory (CTT; Lord & Novick, 1968), factor analysis (FA; Bollen, 1989), and generalizability theory (GT; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). CTT is the oldest of the three approaches (Spearman, 1904, 1910). The FA approach can be argued to represent a refinement of CTT-based reliability, but it can also be argued that the FA conceptualization is different. To clarify these similarities and differences, we discuss the CTT and FA models in one section. GT is conceptually distinct from both CTT and FA, so it is discussed in a separate section. For each approach, we distinguish the model for the measurement value from the method that estimates the parameters or other unknown quantities that are based on the model. For example, CTT defines a model for the measurement value on which a definition of reliability is based, and the well-known coefficient alpha (e.g., Cronbach, 1951) is one of several methods that can be used to estimate CTT reliability.

The motivation for this commentary is that reliability is regularly misunderstood and incorrectly used in research (Cortina, 1993; Schmitt, 1996; Sijtsma, 2009; Sijtsma & Emons, 2011; see also Bentler, 2009; Green & Yang, 2009; Revelle & Zinbarg, 2009). An example is that different methods used to estimate reliability are often confused with different models for the measurement value, suggesting that different methods define different kinds of reliability. For instance, coefficient alpha is sometimes said to represent "internal consistency" reliability, disregarding that alpha is a lower bound to CTT reliability and wrongly suggesting that it represents a reliability definition different from CTT reliability. In addition, researchers often cite misguided

Klaas Sijtsma, PhD, is Professor, Department of Methodology and Statistics, Tilburg University, The Netherlands.
L. Andries van der Ark, PhD, is Associate Professor, Research Institute for Child Development and Education, University of Amsterdam, The Netherlands.
DOI: 10.1097/NNR.0000000000000077

advice that a test is suitable for precise measurement of individuals when a test score's reliability exceeds .80 or .90 (e.g., Nunnally & Bernstein, 1994; Streiner & Norman, 2008) and fail to see that they need to consider the standard error of measurement for that purpose. We cannot go into the whole array of interpretation problems (including those arising from standard error of measurement) associated with reliability but instead discuss the different models and distinguish them from the methods used to estimate reliability in a particular model. We expect such a discussion to provide clarity in the sometimes confusing field of reliability.

In addition to the three mainstream approaches, reliability methods have been proposed for ordinal scores (Schulman & Haden, 1975) and for the estimated latent variable in the Rasch model, where the method is known as the index of subject separation (Andrich, 1982; Gustafsson, 1977). As far as we are aware, neither approach has attained foothold. Alternatively, item response theory focuses on use of Fisher information—which implies scale-dependent information on measurement precision (van der Linden & Hambleton, 1997). We consider these and other options (e.g., Bartholomew & Schuessler, 1991), however interesting, beyond the present discussion. In what follows, we discuss the CTT, FA, and GT approaches to reliability, followed by a computational example that illustrates the three approaches.

THREE APPROACHES TO RELIABILITY

CTT and FA

The Models We assume that a test consists of J items with scores X_j ($j = 1,\ldots, J$). The sum of the item scores produces the test score; we use the notation $X_+ = \sum_{j=1}^{J} X_j$ to indicate that the test score X_+ is the sum of the j item scores. CTT assumes that any observable measurement value Y, such as a psychological test score $Y = X_+$ or an item score $Y = X_j$, consists of a systematic or reliable component, denoted true score T, and a random component, denoted measurement error E. The theory applies as well to temperature estimated by the visual comparison of the top of a mercury column with a scale and the time an athlete running the 200-meter dash produces. The model is

$$Y = T_Y + E_Y. \quad (1)$$

We further focus on test scores $Y = X_+$ and item scores $Y = X_j$.

For any measurement value, CTT leaves the composition of the true score unspecified. For an item score $Y = X_j$, the model is

$$X_j = T_j + E_j. \quad (2)$$

The true score for an individual, indexed i, on item j is operationalized in CTT as $T_{ij} = \varepsilon(X_{ij})$, where the expectation ε is taken across an infinite number of hypothetical, independent replications of the same measurement procedure. This is essentially a thought experiment involving the same item j that is part of a test and administered over and over again to the same individual i, with independence conceived to result from "brainwashing" the measured person between test administrations. This treatment is impossible in practice; hence, the procedure is hypothetical, and the distribution of item scores for person i is unknown (Lord & Novick, 1968, p. 29). For individual i, measurement errors E_{ij} on different replications are random, and across persons, measurement errors obtained in one administration covary 0 with any other variable Z in which E_j is not included; that is, denoting the covariance by $\sigma(.,.)$ CTT assumes that $\sigma(E_j, Z) = 0$. In CTT, the idea is that a measurement value is liable to random influences and that what one really would like to know is the mean performance across a large number of repetitions of the same measurement procedure so as to eliminate random error.

The true scores of different items j and k, which are simply expected values of observable item scores, need not have something in common; that is, T_j and T_k are systematic parts of the two item scores, but there is nothing in the formulation of CTT that necessarily ties the two true scores together in a substantive way. Hence, CTT does not impose restrictions on the association between item scores. CTT as a measurement theory does not reflect the intention test constructors have that the different items in a test should measure the same attribute by means of different operationalizations—items—of the attribute, such as different aspects of pain or social functioning. The model only decomposes a measurement value into a systematic part and a random measurement part.

FA decomposes each of the J item scores X_j that together produce a test score as a weighted linear combination of at most M latent variables or factors (henceforth, latent variables) denoted ξ_m ($m = 1,\ldots,M$; $M \leq J$), such that

$$X_j = b_j + \sum_{m=1}^{M} a_{jm}\xi_m + \delta_j, \quad (3)$$

with intercept b_j, latent-variable loadings α_{jm}, and residual δ_j. All parameters are dependent on item j. The residual δ_j is the sum of a specific or unique (henceforth, unique) error U_j, which represents the systematic part of the item score, and a random error E_j (Bollen, 1989, p. 219) as in CTT, so that $\delta_j = U_j + E_j$ is the part of the item score that in regression terminology is not linearly predicted by latent variables $\xi_1,\ldots,\xi_M$. The unique error component U_j is a composite of all the sources that systematically influence the responses to only item j but not the responses to the other J-1 items in the test. (These systematic sources are usually unknown.) Unique error U_j covaries 0 with latent variable ξ and random error E_j. Relating CTT to FA yields $T_j = b_j + \sum a_{jm}\xi_m + U_j$.

Latent variables $\xi_1,\ldots,\xi_M$ are quantities the different items in the test have in common and that produce correlations between the items. This is a principal difference with CTT, where the true score depends on the item and different items may measure unrelated attributes or unrelated sets of attributes, thus leaving correlations between items dependent on shared causes driving responses unspecified by the model.

Comparison of the CTT and FA Models Whether CTT and FA define different models for the item scores X_j (and other measures) depends on one's perspective. Mathematically, one might say that CTT does not model true score T and that FA models the true score by introducing latent variables and a unique variable that drive responses to item j. Conceptually, there is a principled difference between CTT and FA. CTT does not ask about the origin of the item response and, thus, sets the stage for a reliability definition that quantifies the degree to which one would obtain the same set of measures in a group if one could retest the group under the same circumstances. This reliability definition ignores what each of the items measures and, thus, separates reliability from validity. FA introduces explanatory variables (the latent variables) and, thus, the issue of what the items measure in common (e.g., Revelle & Zinbarg, 2009); that is, FA introduces validity in the model for T.

By introducing validity, two issues become relevant. The first issue is whether the hypothesized FA model fits the data well enough to serve as an explanatory model for the item scores. The second issue is whether the latent variables are correct conceptualizations of the attributes the test intends to measure. For example, if the items are intended and indeed found to have one latent variable in common (Bollen, 1989, pp. 218–221; Bollen leaves out the intercept b_j), they follow the model

$$X_j = b_j + a_j\xi + \delta_j. \tag{4}$$

However, this does not by itself imply that the latent variable ξ is a meaningful representation of the attribute the researchers expect the test to measure. Additional studies have to clarify this representation issue—which is part of the validity research. For example, a set of items intended to measure general health perceptions may also inspire respondents to let moral issues about responsible health behavior enter the response process. Moreover, the items may use complex wording, letting language skills be another response determinant. Hence, latent variable ξ may represent an amalgamation of three sources that affect responding, thus showing that the introduction of validity at least seems to complicate the measurement model.

CTT: The Method

We now concentrate on test scores, so that $Y = X_+$, and discuss a few standard results of CTT; see Lord and Novick (1968) for details. We focus on estimating the quantity derived from the CTT model that is the topic of this article, which is reliability.

Definition of Reliability Let $\sigma^2_{(.)}$ denote the variance. CTT formally defines the properties of test scores a group obtains upon repeated administration of the same test under the same circumstances. Suppose one collects two sets of test scores. Let the test scores be denoted X_+ and X_+', and assume that they have the next two properties:

- for each tested individual, $T_i = T_i'$; and
- in the group, $\sigma^2_{X_+} = \sigma^2_{X_+'}$.

Test scores having these two properties are called parallel. Except for the random measurement error in each test score, two parallel test scores have exactly the same psychometric properties.

Reliability is defined as the product–moment correlation between two parallel test scores, X_+ and X_+', and is denoted $\rho^{CTT}_{X_+X_+'}$. It can be shown that, under the assumptions of CTT, this correlation equals the proportion of true score variance in either test score X_+ or X_+', and that these proportions are equal:

$$\rho^{CTT}_{X_+X_+'} = \frac{\sigma^2_T}{\sigma^2_{X_+}} = \frac{\sigma^2_{T'}}{\sigma^2_{X_+'}}. \tag{5}$$

Reliability is a proportion; hence, its values lie between 0 and 1. Value 0 means that all individuals have the same true score, meaning that the test does not distinguish individuals. Value 1 means that all test score variance equals true score variance. What this means for measurement error can be seen if one rewrites reliability, for example, for test score X_+. We use the CTT result that $\sigma^2_{X_+} = \sigma^2_T + \sigma^2_E$. Then, reliability can be written as

$$\rho^{CTT}_{X_+X_+'} = 1 - \frac{\sigma^2_E}{\sigma^2_{X_+}}. \tag{6}$$

Equation 6 shows that reliability equal to 1 implies that $\sigma^2_E = 0$; thus, each tested person has the same measurement error. Lord and Novick (1968, pp. 36–37) show that the mean measurement error equals 0, that is, $\mu_E = 0$. Hence, perfect reliability implies that each person is measured error free, that is, $E_i = 0$ and $X_{+i} = T_i$. In practical measurement, reliability values are often between .60 and .95, and values in excess of .80 or .90—depending on the research context—are often deemed sufficiently high for the measurement purpose envisaged.

Finally, reliability can be shown to equal the squared correlation between the test score and the true score:

$$\rho^{CTT}_{X_+X_+'} = \rho^2_{X_+T}. \tag{7}$$

In a linear regression context, reliability equals the proportion of variance in X_+ explained by T. Hence, reliability is a group characteristic and not an indicator of the measurement

precision of individual test scores; see Mellenbergh (1996) for a discussion of the difference between reliability and measurement precision.

Estimation of Reliability The CTT model in Equation 1 is an equation with two unknowns, T and E; hence, it is a tautological model. Consequently, true score variance σ_T^2 (Equation 5), measurement-error variance σ_E^2 (Equation 6), and correlation ρ_{X_+T} (Equation 7) are unobservable, and reliability cannot be estimated directly from the data. Several alternative estimation methods have been proposed. Without going into detail, we mention the following:

- Parallel-test reliability: Two test forms are constructed that approximate parallelism as close as possible, and the test scores are correlated. Because real tests are never truly parallel, for real data, the correlation underestimates the true reliability.
- Retest reliability: The same test is readministered after a suitable amount of time, and the correlation between both administrations estimates reliability. Experience shows that, usually, this estimate is much lower than other reliability estimates (Sijtsma, 2012).
- Single-test, single-administration estimates:
  - Split-half reliability: The correlation between two half tests estimates the reliability of a test half the length of the test of interest, and the Spearman–Brown prophesy formula can be used to estimate reliability for the whole test. This method produces either underestimates or overestimates of reliability (van der Ark, van der Palm, & Sijtsma, 2011).
  - Covariance-based estimates: Estimates based on the covariances of all item pairs provide lower bounds to the reliability. Coefficient alpha (Cronbach, 1951; Guttman, 1945; speaks of coefficient lambda3) is the best known representative of this category. Coefficient lambda2 (Guttman, 1945) and the greatest lower bound are other representatives (Bentler & Woodward, 1980; Ten Berge & Sočan, 2004).

FA: The Method

Definition of Reliability FA defines reliability for X_+ adopting the CTT definition; see Mellenbergh (1994, 1998) for an alternative one-factor model without a unique error component U_j, but including random measurement error E_j in which reliability for a single, estimated latent variable $\hat{\xi}$ rather than test score X_+ is defined, and a method is discussed to estimate the reliability of $\hat{\xi}$.

We take the multifactor model in Equation 3 as a point of departure. We assume that each item loads on at least one latent variable and that at least two items load on the same latent variable. These restrictions rule out unique components of which the variance $\sigma_{U_j}^2$ cannot be estimated anyway (Bollen, 1989, pp. 220–221), thus producing a simpler model. Adopting multiple latent variables to some extent may compensate for the absence of unique variables, but we do not include in the model observable explanatory variables that may influence the item scores as Bollen (1989, pp. 220–221) suggested, thus limiting attention to attributes represented by latent variables that affect test performance. A further simplification is obtained by assuming that, for all J items, intercept $b_j = 0$. Next, we write test score X_+ as the sum of the item scores:

$$X_+ = \sum_{j=1}^{J} X_j = \sum_{j=1}^{J} \sum_{m=1}^{M} a_{jm}\xi_m + \sum_{j=1}^{J} E_j, \tag{8}$$

with, at the test level, $E = \sum_j E_j$, and true score T defined as

$$T = \sum_{j=1}^{J} \sum_{m=1}^{M} a_{jm}\xi_m. \tag{9}$$

Following Raykov and Shrout (2002), CTT reliability of test score X_+ can be written as

$$\rho_{X_+X_+'}^{FA} = \frac{\sigma_T^2}{\sigma_{X_+}^2} = \frac{\sigma^2\left(\sum_j \sum_m a_{jm}\xi_m\right)}{\sigma^2\left(\sum_j \sum_m a_{jm}\xi_m + \sigma_E^2\right)}. \tag{10}$$

This reliability definition is known as coefficient ω (McDonald, 1999; Revelle & Zinbarg, 2009).

Estimation of Reliability The right-hand side in Equation 10 has been equated to CTT reliability, thus assuming that the FA model with M latent variables ξ_m fits the data sufficiently well to make the equality correct by approximation. In other words, the FA model the researcher hypothesized has to be the correct model for item performance; either forgetting important latent variables or including irrelevant latent variables, misspecifying the loading structure, or a combination of these produces misfit of the model to the data and an inequality in Equation 10.

Raykov and Shrout (2002) discussed a structural equation modeling approach to estimating coefficient ω. Revelle and Zinbarg (2009) discussed Equation 3 (their Equation 16, which is basically the same equation as Equation 3) for one common factor and several group factors, that is, factors on which subsets of items load but not all items, and rewrote Equation 10 accordingly. Distinguishing a common factor and group factors does not produce another reliability coefficient than Equation 10, but it does hypothesize another latent variable structure for the items. The latent variable structure's fit to the data has to be investigated before one can say that the resulting coefficient ω equals CTT reliability.

Revelle and Zinbarg (2009) rephrased Equation 10 using an item's communality, denoted h_j^2, which is the proportion of variance of the item score explained by the common latent variable and the group variables, say, M latent variables in total, so that $h_j^2 = \sum_{m=1}^{M} a_{jm}^2$. The complement, $d_j^2 = 1 - h_j^2$, equals the proportion of unexplained variance and can be considered an approximation of the item's proportion of error variance. Using these proportions for standardized item scores, coefficient ω

becomes coefficient ω_t (McDonald, 1999; Revelle & Zinbarg, 2009):

$$\omega_t = 1 - \frac{\sum_j \left(1 - b_j^2\right)}{\sigma_{X_+}^2} = 1 - \frac{\sum_j d_j^2}{\sigma_{X_+}^2}. \quad (11)$$

Another method called coefficient ω_h assesses how well the test measures the common latent variable (Revelle & Zinbarg, 2009) and simplifies Equation 11 by letting $M = 1$.

GT: Model and Method

GT's purpose is to identify as many sources of error as possible that affect relative or absolute interpretations of persons' test scores and then correct the reliability coefficient taking the error sources into account. Relative interpretation refers to a person's position with respect to other tested persons, and absolute interpretation refers to a person's position relative to an absolute performance criterion—independent of other persons' test scores (Brennan, 2001, p. 13). The reliability coefficient is labeled generalizability coefficient for relative interpretations and index of dependability for absolute interpretations. The error correction should provide the researcher with an impression of the degree to which persons' test scores can be generalized accurately to their mean test score they would acquire under all possible conditions the test user is ready to accept. An example is different test versions randomly sampled from the same set of all possible items taking a particular attribute. Reliability is then corrected taking the interaction effect involving persons and test forms suggesting persons are differently ordered by different tests into account.

To realize its goal, in a generalizability study, GT uses the statistical framework of analysis of variance (ANOVA) to decompose test-score variance into as many sources as the test design allows, after which a decision study uses this information to devise a final, efficient assessment procedure that avoids as many error sources as possible. We only discuss the basics of the generalizability study that allow us to define a generalizability coefficient.

Model. The simplest GT model considers a population of persons that respond to, say, the multiple-choice items from the set of all possible multiple-choice items for a particular attribute. The set of items constitutes the universe of observations. In practice, one considers a sample of persons and a sample of items, which constitute the test. In ANOVA parlance, this is the fully crossed, one-facet random effects design (e.g., Brennan, 2001; Sanders, 1998, 2005; Shavelson & Webb, 1991). Facets are in GT what factors are in ANOVA (but not in FA). An individual's test score is decomposed following the ANOVA layout of a dependent variable.

Let $\mu = \varepsilon_i \varepsilon_j (X_{ij})$ denote the grand item-score mean across the population of persons and the set of items; $\mu_i = \varepsilon_j (X_{ij})$ denote the mean person score across all items in the item set, known as the universe score; and $\mu_j = \varepsilon_i (X_{ij})$ denote the mean item score across all persons in the population. In addition, let μ_{ij} denote the mean score for the combination of person i and item j, for example, across a number of different raters in a design that includes raters as a second facet, to be discussed shortly. It proves convenient to define the following effects (Brennan, 2001, p. 63): $\nu_i = \mu_i - \mu$ is the person effect, $\nu_j = \mu_j - \mu$ is the item effect, and $\nu_{ij} = \mu_{ij} - \mu_i - \mu_j + \mu$ is the interaction effect.

Analogous to CTT, observed score X_{ij} is decomposed in a structural part and random measurement error,

$$X_{ij} = \mu_{ij} + E_{ij}, \quad (12)$$

in which

$$\mu_{ij} = \mu + \nu_i + \nu_j + \nu_{ij}, \quad (13)$$

so that

$$X_{ij} = \mu + \nu_i + \nu_j + \nu_{ij} + E_{ij}. \quad (14)$$

Because, in the fully crossed, one-facet random effects design, only one score X_{ij} is available for the combination of a person and an item, mean μ_{ij} cannot be estimated. Hence, interaction effect ν_{ij} and random measurement error E_{ij} cannot be distinguished and are combined to constitute the residual part of the model, defined as $\Delta_{ij} = \nu_{ij} + E_{ij}$. The residual, thus, confounds the interaction component and the error component, with the latter containing all remaining influences on the item score that are not in Equation 14. The decomposition of X_{ij} that is used in GT equals

$$X_{ij} = \mu + \nu_i + \nu_j + \Delta_{ij}. \quad (15)$$

The GT model can be extended, for example, by including raters. Items and raters, for example, also sampled from the set of all possible raters, together constitute the universe of observations. Let raters be indexed r ($r = 1,\ldots,R$), and let X_{ijr} be the score on item j that rater r assigned to person i. Let μ_{ijr} be the expected score that is decomposed into a person effect, an item effect, a rater effect, and interaction effects, such that

$$\mu_{ijr} = \mu + \nu_i + \nu_j + \nu_r + \nu_{ij} + \nu_{ir} + \nu_{jr} + \nu_{ijr} \quad (16)$$

and

$$X_{ijr} = \mu + \nu_i + \nu_j + \nu_r + \nu_{ij} + \nu_{ir} + \nu_{jr} + \Delta_{ijr}. \quad (17)$$

This extension shows the potential of GT for use in many important measurement situations in health and nursing research.

Definition and Estimation of Generalizability GT's reliability coefficient, known as the generalizability coefficient, equals the proportion of person-score variance relative to all effects that influence relative person ordering, thus excluding all effects the relative person ordering does not depend on and that define the scope of the generalization. We consider the situation in which the test is a randomly drawn set of J items from the set of all possible items for measurement of the attribute of interest; see Shavelson and Webb (1991) and Brennan (2001) for details. We discuss the generalizability coefficient for relative person effects (rpe's), defined as $\nu_i = \mu_i - \mu$. Let $\sigma_i^2 = \varepsilon_i(\nu_i)^2$ equal the variance of the rpe's in the group, let $\sigma_j^2/J = \varepsilon_j(\nu_j)^2/J$ denote the variance of the mean item score across randomly drawn tests, and let $\sigma_\Delta^2/J = \varepsilon_i\varepsilon_j(\nu_{ij}+E_{ij})^2/J$ denote the residual variance for mean item scores; then, in the population, the mean item-score variance can be shown to be equal to

$$\sigma_X^2 = \sigma_i^2 + \frac{\sigma_j^2}{J} + \frac{\sigma_\Delta^2}{J}. \quad (18)$$

The generalizability coefficient for rpe's is defined as

$$\rho_{rpe}^{GT} = \frac{\sigma_i^2}{\sigma_i^2 + \frac{\sigma_\Delta^2}{J}}. \quad (19)$$

The denominator contains the variance components that influence the relative standing of persons, which are σ_i^2 and σ_Δ^2/J —the latter containing the interaction of persons and items, which suggests that the standing of persons relative to one another depends on the test. The denominator does not contain the variance of mean item scores across random tests, σ_j^2/J, as random sampling of items does not systematically affect relative person standing. Thus, the denominator is unequal to the test-score variance that appears in the CTT and FA reliability definitions in Equation 5 and Equation 10, respectively. Shavelson and Webb (1991), Sanders (1998, 2005), and Brennan (2001) discuss the estimation of the variance components in Equation 19, following ANOVA logic.

For the computational example to be discussed later, we used a fully crossed, two-facet, random effects design, including person, item, and rater main effects; three two-way interaction effects; and a residual effect that combines inseparable three-way interaction and error terms (Equation 17). The mean item-score variance can be decomposed as (Brennan, 2001, p. 11)

$$\sigma_X^2 = \sigma_i^2 + \frac{\sigma_j^2}{J} + \frac{\sigma_r^2}{R} + \frac{\sigma_{ij}^2}{J} + \frac{\sigma_{ir}^2}{R} + \frac{\sigma_{jr}^2}{JR} + \frac{\sigma_\Delta^2}{JR}. \quad (20)$$

The generalizability coefficient is defined as

$$\rho_{rpe}^{GT} = \frac{\sigma_i^2}{\sigma_i^2 + \frac{\sigma_{ij}^2}{J} + \frac{\sigma_{ir}^2}{R} + \frac{\sigma_\Delta^2}{JR}}. \quad (21)$$

The denominator contains the variance terms that influence relative person standing, which are the interaction terms for persons and tests and persons and raters (in addition to σ_i^2). More complex designs render definitions of the generalizability coefficient that differ with respect to the variance components in the denominators.

ADDITIONAL REMARKS

In this section, we briefly discuss topics that are relevant for the discussion on reliability.

Comparison of FA and GT

The FA approach decomposes the test score into common and unique latent variables that are hypothesized to relate to attributes driving item responses. The GT approach decomposes the test score into effects because of person performance but also item difficulty, rater level, time of testing, and so on. Rather than introducing the meaning of measurement into the equation, GT asks to which degree test scores are dependent on item difficulty, rater level, time of testing, and so on and, thus, approaches the usefulness of the test score from a utilitarian perspective. Depending on the design of the study, test-score reliability—here, generalizability—is defined differently as different variance components enter the generalizability coefficient. Interestingly, this also means that the test-score variance does not enter the denominator of the generalizability coefficient; instead, the variance components that are relevant for person variance are in the denominator.

Model Choice and Goodness of Fit

An FA model serves as a hypothesis for the attribute structure that drives responses to items and determines the interitem correlation structure. The FA model can be tested statistically in a confirmatory FA or structural equation modeling framework. If the FA model does not fit because it is incomplete, latent or manifest (e.g., gender) variables that add to the explanation of the interitem correlation structure, but that were not included in the model, may be added, and the alternative FA model may be tested next. Incompleteness of models—not only FA models—is the common state of affairs; that is, models serve as idealizations of the truth and, thus, at best approximate the truth. If an FA model does not fit the data well but specifies the correct number of latent variables, it may be modified by having a different loading structure in which particular loadings are restricted to be 0, to be equal to other loadings, or to have particular values. Suggestions how to do this may be derived ad hoc from the goodness-of-fit analysis or may be hypothesized

on the basis of theoretical expectations after which the model is tested as the null hypothesis.

Goodness of Fit of CTT?

In CTT, parallelism (defined earlier) and essential tau-equivalence (for a set of tests, a characteristic that defines individuals' true scores to be different by a constant depending on true-score pairs and allows different error variances in the group; see Lord & Novick, 1968, p. 50) imply testable restrictions on statistics (e.g., Graham, 2006; Lord & Novick, 1968). For example, two parallel tests have the same correlations with any other variable Z; that is, $\rho_{X_+Z}=\rho_{X'_+Z}$. Of course, correlations usually will be different for different variables Z_q and Z_t. Likewise, essential tau-equivalence implies testable consequences. One could argue that the empirical investigation of these consequences represents ways of investigating the goodness of fit of CTT to the data. However, this argument is incorrect as the CTT model equation $X = T + E$ is a tautology and does not restrict the data; hence, the fit of CTT to the data cannot be investigated. The investigation of the observable consequences of parallelism and essential tau-equivalence provides evidence of whether parallel-test reliability and methods like coefficient alpha, lambda2, and the greatest lower bound provide underestimates of the reliability.

COMPUTATIONAL EXAMPLE

To illustrate the three types of reliability, we first constructed a statistical simulation model according to the setup of a study described in National Council of State Boards of Nursing (2009). Then, we sampled a large data set from the simulation model and computed the generalizability coefficient (GT; Equation 21), coefficient alpha (CTT; σ_{jk} is the interitem covariance),

$$alpha=\frac{J}{J-1}\,\frac{\sum\sum_{j\neq k}\sigma_{jk}}{\sigma^2_{X_+}} \tag{22}$$

and coefficient ω_t (FA; Equation 11). We used a statistical simulation model so as to have full control over the data generation.

In the National Council of State Boards of Nursing (2009) study, the performances of nursing sciences students in three scenarios—new onset of chest pain, sudden onset of shortness of breath, and acute change in level of consciousness—were videotaped. Two faculty members independently rated each videotape using a multi-item questionnaire with respect to patient–nurse relationship, symptom recognition, assessment, and intervention. Items were scored using rank numbers for the answer categories "inadequate," "somewhat inadequate," "not attempted," "somewhat adequate," and "adequate."

We used a multifactor model (Equation 3) with an additional rater component as the statistical simulation model. We assumed that five abilities represented by five latent variables drove the responses to 28 items (J = 28). The first latent variable represented general ability, and the other four latent variables represented patient–nurse relationship ability, symptom recognition ability, assessment ability, and intervention ability. Let ξ_{im} (i = 1,…, N; m = 1,…,5) denote student i's value on latent variable m. Let b_{jr} (j = 1,…,J; r = 1,…,R) be the intercept of item j when assessed by rater r. A larger intercept means that the item is easier. Subscript r indicates that the rater also influences the item easiness: The item is "easier" if the rater is lenient and "more difficult" if the rater is strict. Let α_{jmr} denote the loading of item j on latent variable m, when assessed by rater r. The loadings represent the weight of latent variable m on item j. Subscript r indicates that the rater also influences the latent variable's loading: If the rater values a particular aspect, such as symptom recognition, more important, the loading is larger. Let E_{ij} denote random error. Let ϑ_{ij} denote a continuous score underlying the item scores.

The model for continuous score ϑ_{ij} is

$$\vartheta_{ij}=b_{jr}+\sum_{m=1}^{5}a_{jmr}\xi_{im}+E_{ij}. \tag{23}$$

Continuous scores ϑ_{ij} were transformed into discrete item scores X_{ij} valued 0 (*inadequate*), 1 (*somewhat inadequate*), 2 (*not attempted*), 3 (*somewhat adequate*), and 4 (*adequate*) as follows. First, ϑ_{ij} was transformed to standard scores, that is, $z_\vartheta=\left(\vartheta_{ij}-\overline{\vartheta}\right)/\hat{\sigma}_\vartheta$; then two points were added, producing a mean item score equal to 2; and finally, the scores were rounded to the nearest admissible value X_{ij}.

Data were generated using R (R Core Team, 2014; code available from the second author). Let $N(\mu, \sigma^2)$ denote a normal distribution with mean μ and variance σ^2. We sampled latent variables for 2,000 simulees from $N(0,1)$. We chose a sample size much larger than the National Council of State Boards of Nursing study (N = 37) to eliminate gross effects from sampling fluctuation. We chose the correlations between the latent variables equal to 0.4, which is a reasonable choice given that the abilities probably are positively correlated. Consistent with the National Council of State Boards of Nursing (2009) study, we assumed that two raters (R = 2) assessed each simulee's videotapes. Intercept b_{jr} is the sum of an item component b^*_j, sampled from $N(0,1)$, and a rater component b^*_r, independently sampled from $N(0,0.2)$; that is, $b_{jr}=b^*_j+b^*_r$. As the distribution of b^*_j has a larger variance than the distribution of b^*_r, the item has a larger effect on the easiness than the rater.

The loadings were obtained as follows. The first ability is a general ability on which all 28 items have a loading equal to 1. Loading α_{j1r} is the sum of an item component a^*_{j1}, sampled from $N(0.5, 0.5)$, and a rater component a^*_{1r}, independently sampled from $N(0, 0.3)$; that is, $a_{j1r}=a^*_{j1}+a^*_{1r}$. The item has a larger effect on the loading than the rater. The other four abilities are domain specific and drive only subgroups of items. Only Items 1–7, which concern patient–nurse relationship, load on

the second latent variable; the loadings of the other items equal 0. Loading α_{j2r} consists of an item component a^*_{j2}, sampled from $N(0,0.5)$, and a rater component a^*_{2r}, independently sampled from $N(0, 0.2)$. Hence, $a_{j2r}=a^*_{j2}+a^*_{2r}$ for $j = 1,\ldots, 7$, and $a_{j2r} = 0$, otherwise. Following the same line of reasoning, $a_{j3r}=a^*_{j3}+a^*_{3r}$ for $j = 8,\ldots, 14$, and $a_{j3r} = 0$, otherwise; $a_{j4r}=a^*_{j4}+a^*_{4r}$ for $j = 15,\ldots, 21$, and $a_{j4r} = 0$, otherwise; and $a_{j5r}=a^*_{j5}+a^*_{5r}$ for $j = 22,\ldots,28$, and $a_{j5r} = 0$, otherwise. Random error E_{ij} was sampled from $N(0,2)$.

Because of the three-way structure of the simulated data (students × items × raters), one should estimate generalizability rather than reliability. We advise against using CTT reliability estimators (cf. National Council of State Boards of Nursing, 2009) or FA estimators because, by ignoring that each student appears twice in the data, one incorrectly treats the observations as independent; hence, estimators, such as alpha and ω_t, that ignore the dependence structure in the data cannot be interpreted meaningfully.

The generalizability coefficient was computed using our own R code (available from the second author). We found that $\hat{\sigma}^2_i = .0398$, $\hat{\sigma}^2_{ij} = .0920$, $\hat{\sigma}^2_{ir} = .0187$, and $\hat{\sigma}^2_{\Delta} = .7160$; hence, the generalizability coefficient for rpe's (Equation 21) equals

$$\hat{\rho}^{GT}_{rpe} = \frac{\hat{\sigma}^2_i}{\hat{\sigma}^2_i + \frac{\hat{\sigma}^2_{ij}}{J} + \frac{\hat{\sigma}^2_{ir}}{R} + \frac{\hat{\sigma}^2_{\Delta}}{JR}}$$

$$= \frac{.0398}{.0398 + \frac{.0920}{28} + \frac{.0187}{2} + \frac{.7160}{56}} = .6102. \quad (24)$$

The estimated alpha and ω_t coefficients (computed using the R package psych, version 1.3.2; Revelle, 2014) equaled .6698 and .7470, respectively, but should not be used for these data.

The CTT approach and the FA approach should only be used for two-way data. For example, coefficients alpha and ω_t for the data Rater 1 produced equal .3569 and .6575, respectively, and for the data Rater 2 produced, the values are .7788, and .8165, respectively. Sirotnik (1970) showed that, for two-way data (e.g., persons × items), the generalizability coefficient equals coefficient alpha. The results show that Rater 2 produced more reliable test scores than Rater 1. Coefficient ω_t exceeds alpha. For these data sets, it can be expected that, compared with coefficient alpha, coefficient ω_t is a better approximation to the reliability because it takes the simulated factor structure into account.

Accepted for publication December 8, 2014.

The authors thank Susan Henly and Piet Sanders for their comments on a previous draft of this article. Of course, views presented here and possible errors are the authors' responsibility.

The authors declare no conflicts of interest.

Corresponding author: Klaas Sijtsma, PhD, Department of Methodology and Statistics, TSB, Tilburg University, PO Box 90153, 5000LE Tilburg, The Netherlands (e-mail: k.sijtsma@tilburguniversity.edu).

REFERENCES

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Rasch Models for Measurement in Educational and Psychological Research. Education Research and Perspectives*, *9*, 95–104. Retrieved from http://www.rasch.org/erp7.htm

Bartholomew, D. J., & Schuessler, K. F. (1991). Reliability of attitude scores based on a latent trait model. *Sociological Methodology*, *21*, 97–123.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*, 137–143. doi:10.1007/s11336-008-9100-1

Bentler, P. M., & Woodward, J. A. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. *Psychometrika*, *45*, 249–267. doi:10.1007/BF02294079

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104. doi:10.1037/0021-9010.78.1.98

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. doi:10.1007/BF02310555

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

de Jong, A. E. E., Bremer, M., Schouten, M., Tuinebreijer, W. E., & Faber, A. W. (2005). Reliability and validity of the pain observation scale for young children and the visual analogue scale in children with burns. *Burns*, *31*, 198–204. doi:10.1016/j.burns.2004.09.013

Gandek, B., Sinclair, S. J., Kosinski, M., & Ware, J. E. Jr. (2004). Psychometric evaluation of the SF-36 health survey in Medicare managed care. *Healthcare Financing Review*, *25*, 5–25. Retrieved from http://www.cms.gov/Research-Statistics-Data-and-Systems/Research/HealthCareFinancingReview/downloads/04summerpg5.pdf

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. doi:10.1177/0013164406288165

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*, 121–135. doi:10.1007/s11336-008-9098-4

Gustafsson, J.-E. (1977). *The Rasch model for dichotomous items: Theory, applications and a computer program*. Technical report no. 63, Institute of Education, University of Göteborg, Sweden.

Guttman, L. (1945). A basis for analyzing test–retest reliability. *Psychometrika*, *10*, 255–282. doi:10.1007/BF02288892

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73. doi:10.1111/jedm.12000

Lissitz, R. W. (Ed.). (2009). *The concept of validity. Revisions, new directions, and applications*. Charlotte, NC: Information Age Publishing.

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

Ma, C., Chen, S., You, L., Luo, Z., & Xing, C. (2012). Development and psychometric evaluation of the Treatment Adherence Questionnaire of patients with hypertension. *Journal of Advanced Nursing*, *68*, 1402–1413. doi:10.1111/j.1365-2648.2011.05835.x

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

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

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*, 223–236. doi:10.1207/s15327906mbr2903_2

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*, 293–299. doi:10.1037/1082-989X.1.3.293

Mellenbergh, G. J. (1998). Het één-factor model voor continue en metrische responsen. [The one-factor model for continuous and metric responses]. In W. P. van den Brink, & G. J. Mellenbergh (Eds.), *Testleer en testconstructie* (pp. 155–186). Amsterdam, The Netherlands: Boom.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.

National Council of State Boards of Nursing (2009). *Report of findings from "The effect of high-fidelity simulation on nursing students' knowledge and performance: A pilot study."* (Research Brief Vol. 40). Chicago, IL: Author. Retrieved from https://www.ncsbn.org/09_SimulationStudy_Vol40_web_with_cover.pdf

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

R Core Team (2014). *R: A language and environment for statistical computing [computer programming language]*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling*, *9*, 195–212. doi:10.1207/S15328007SEM0902_3

Revelle, W. (2014). *psych: Procedures for personality and psychological research.* R package version 1.4.2. Retrieved from http://personality-project.org/ and www.cran.r-project.org/web/packages/psych/psych.pdf

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the GLB: Comments on Sijtsma. *Psychometrika*, *74*, 145–154. doi:10.1007/s11336-008-9102-z

Sanders, P. F. (1998). Generaliseerbaarheidstheorie [Generalizability theory]. In W. P. van den Brink & G. J. Mellenbergh (Eds.), *Testleer en testconstructie* [Test theory and test construction] (pp. 95–131). Amsterdam, The Netherlands: Boom.

Sanders, P. F. (2005). Generalizability theory: Estimation. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 711–717). Chichester, UK: Wiley.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350–353. doi:10.1037/1040-3590.8.4.350

Schulman, R. S., & Haden, R. L. (1975). A test theory model for ordinal measurement. *Psychometrika*, *40*, 455–472. doi:10.1007/BF02291549

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120. doi:10.1007/s11336-008-9101-0

Sijtsma, K. (2012). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika*, *77*, 4–20. doi:10.1007/s11336-011-9242-4

Sijtsma, K. (in press). Classical test theory. In S. J. Henly (Ed.), *Routledge international handbook of advanced quantitative methods in nursing research*. Abingdon, UK: Routledge.

Sijtsma, K., & Emons, W. H. M. (2011). Advice on total-score reliability issues in psychosomatic measurement. *Journal of Psychosomatic Research*, *70*, 565–572. doi:10.1016/j.jpsychores.2010.11.002

Sirotnik, K. (1970). An analysis of variance framework for matrix sampling. *Educational and Psychological Measurement*, *30*, 891–908. doi:10.1177/001316447003000410

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology*, *15*, 201–293. doi:10.2307/1412107

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271–295. doi:10.1111/j.2044-8295.1910.tb00206.x

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: A practical guide to their development and use* (4th ed.). New York, NY: Oxford University Press.

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*, 613–625. doi:10.1007/BF02289858

van der Ark, L. A., van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement*, *35*, 380–392. doi:10.1177/0146621610392911

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.

Wainer, H., & Braun, H. I. (Eds.). (1988). *Test validity*. Hillsdale, NJ: Erlbaum.