Article

On Misconceptions and the Limited Usefulness of Ordinal Alpha

Educational and Psychological Measurement
2018, Vol. 78(6) 1056–1071

DOI: 10.1177/0013164417727036
journals.sagepub.com/home/epm
SAGE

R. Philip Chalmers[1]

Abstract

This article discusses the theoretical and practical contributions of Zumbo, Gadermann, and Zeisser's family of ordinal reliability statistics. Implications, interpretation, recommendations, and practical applications regarding their ordinal measures, particularly ordinal alpha, are discussed. General misconceptions relating to this family of ordinal reliability statistics are highlighted, and arguments for interpreting ordinal alpha as a measure of hypothetical reliability, as opposed to observed reliability, are presented. It is concluded that ordinal alpha should not be used in routine reliability analyses and reports, and instead should be understood as hypothetical tool, similar to the Spearman–Brown prophecy formula, for theoretically increasing the number of ordinal categorical response options in future applied testing applications.

Keywords

ordinal alpha, coefficient alpha, reliability, polychoric correlations, Spearman–Brown prophecy formula

This article discusses the theoretical contributions offered by Zumbo, Gadermann, and Zeisser's (2007) family of statistics for quantifying test reliability. At the time of writing, Zumbo et al.'s article, as well as the follow-up articles by Oliden and Zumbo (2008; written in Spanish) and Gadermann, Guhn, and Zumbo (2012), have jointly received more than 1,100 citations, according to Google Scholar. To put this into perspective with regard to influential bodies of work in psychometrics, according to Google Scholar, Lord and Novick's (1968) seminal book on psychometric test theory has been cited just over 8,000 times, and the classical psychometrics textbook written

[1]The University of Georgia, Athens, GA, USA

Corresponding Author:
R. Philip Chalmers, Department of Educational Psychology, The University of Georgia, 323 Aderhold Hall, Athens, GA 30602, USA.
Email: rphilip.chalmers@uga.edu

by Crocker and Algina (1986) has received around 5,500 citations. Given that the ordinal reliability articles are much newer than these classical psychometric texts, the total citation count is only expected to rise as researchers continue to adopt ordinal reliability in their applied work.

Evidently, the impact of the ordinal reliability statistics in the research community is nontrivial. As well, researchers are generally citing the authors of ordinal reliability for their theoretical and practical contributions to the area of reliability theory (e.g., McNeish, 2018). However, the implications regarding what the ordinal measures of reliability mean to applied researchers, as well as when they should be used in practice, have yet to be meaningfully discussed. Therefore, the purpose of this article is to explain what ordinal α is, why and when it might be useful, and to clarify several misconceptions regarding this family of reliability statistics that appeared in the original articles.

The following presentation focuses exclusively on coefficient α and ordinal α for simplicity; however, the arguments generalize to other forms of ordinal reliability discussed by Zumbo et al. (2007), Oliden and Zumbo (2008), and Gadermann et al. (2012). The article begins by reviewing the concepts from classical test theory so that coefficient α, as well as the newer ordinal α, can be understood. Next, four important misconceptions regarding ordinal α that appeared in the aforementioned authors' articles are presented and discussed at length. The article closes with one potentially useful application for ordinal α, based on a hypothetical mental exercise, that tests analysts may find interesting in their measurement applications.

Classical Test Theory and Reliability

Classical test theory (CTT) is centered on decomposing observed scores into two unobserved components based on observable summary statistics (Lord & Novick, 1968; Traub, 1997). In particular, CTT focuses on the unweighted sum-score statistic as the observed composite variable with which test analysts wish to understand and draw inferences. Let $X_{i1}, X_{i2}, \ldots, X_{in}$ represent the observed item-level scores for individual i given n distinct items in a given measurement instrument, and let $X_i = \sum_{j=1}^{n} X_{ij}$ represent the unweighted total score for person i. The relationship

$$X_i = T_i + E_i \tag{1}$$

is then conceptualized, where the random variable X_i is decomposed into a fixed variable T_i, known as the true score, and random error component E_i, known as the error or residual. By construction, the expected value of E is 0, while the relationship $\mathrm{COV}(T, E) = 0$ is assumed. This simple expression and assumptions represents the cornerstone of nearly all ancillary developments in CTT research (Traub, 1997). Of particular importance in this article is how these ideas can be used to express the *reliability* of the observed test scores X.

Reliability, as a theory of measurement precision, is a methodology relating to how *unobserved* (or latent) variables can be measured using only *observed* data components. The question to be answered is, ''Given the observed information provided

by a test, what can be said about the unobserved reality?'' In the case of ability measurement, for example, we might be interested in an individual's unobserved ''proficiency in mathematics'' given their observed score on a valid test. Reliability is then used to quantify how effectively the test information reflects the precision of the unobserved true scores, thereby providing statistical mechanisms for creating estimates of precision (e.g., the standard error of measurement and large-sample confidence intervals; Lord & Novick, 1968). Note that the field of factor analysis (McDonald, 1985) and latent variable modeling more generally (Bartholomew, Knott, & Moustaki, 2011) have precisely the same purpose, whereby observed data components are studied so that inferences can be drawn about the structure and properties of unobserved variables (commonly referred to as factors or latent traits).

More technically, reliability for the observed composite variable X, in the form of internal consistency (Crocker & Algina, 1986), is formally expressed as

$$r(X) = \mathrm{VAR}(T)/\mathrm{VAR}(X) = 1 - \mathrm{VAR}(E)/\mathrm{VAR}(X). \tag{2}$$

This reliability coefficient is bounded within $[0, 1]$, where values closer to 1 indicate better measurement precision (i.e., less sampling and measurement error). In practice, reliability is estimated using one of several possible estimators, such as coefficient α (Cronbach, 1951; Guttman, 1945), McDonald's ω (McDonald, 1999), Revelle's β (Revelle, 1979), and so on. Here we will focus on coefficient α because of its popularity and its connection to the ordinal measures proposed by Zumbo et al. (2007), both of which are presented below.

Coefficient α

Perhaps the most popular measure of the internal reliability, which was introduced by Guttman (1945) and popularized by Cronbach (1951), is coefficient α. Coefficient α can be expressed using matrix notation as

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\mathrm{tr}\,(\boldsymbol{\Sigma})}{1'\boldsymbol{\Sigma}1}\right), \tag{3}$$

where $\boldsymbol{\Sigma}$ is the $n \times n$ variance–covariance matrix among the test items, and $\mathrm{tr}\,(\boldsymbol{\Sigma})$ is the trace of the matrix. In practice, a sample estimate of the covariance matrix $\hat{\boldsymbol{\Sigma}}$ is used in place of $\boldsymbol{\Sigma}$ to obtain the sample estimate of α. This measure assumes that the items are tau-equivalent, meaning that all items are of equal importance when measuring the unobserved construct, but that the respective error variances for each item are allowed to differ (Lord & Novick, 1968). When $\boldsymbol{\Sigma}$ is replaced by the correlation matrix $\mathbf{R}$, or alternatively every test item is first standardized to have a mean of zero and variance of 1 before computing $\hat{\boldsymbol{\Sigma}}$, then the standardized α estimate will be obtained. Standardized α further assumes that the variance of each item are equal, resulting in the more stringent, and less realistic, assumption that the items are parallel (Lord & Novick, 1968; McDonald, 1985).

An alternative definition for coefficient α based on a factor analytic model with uncorrelated residuals was realized by McDonald (1985), which is expressed as

$$\alpha = \frac{n}{n-1}\left(1 - \frac{n \cdot (\bar{f})^2 - \bar{f}^2}{n \cdot (\bar{f})^2 + \bar{u}^2}\right). \quad (4)$$

In this expression, $\bar{f}$ represents the average of the n factor loadings, $\bar{f}^2$ is the average of the squared factor loadings, and $\bar{u}^2$ is the average of the associated uniqueness terms. Equation 4 is interesting in that it also only requires the items to be tau-equivalent, and because it can be computed using either the covariance or correlation matrix with factor analysis or structural equation modeling software (e.g., Rosseel, 2012). Note that if there is interest in the respective sampling variability of these factor loadings then the covariance matrix should be supplied to the software package instead of the correlation matrix to avoid estimation issues relating to the bounded nature of correlations (Cudeck, 1989).

Zumbo et al.'s Ordinal α

Zumbo et al. (2007) proposed a new reliability statistic, which they termed ''ordinal α,'' that was created to allegedly account for the categorical nature of the item response stimuli commonly found in educational tests, psychological surveys, clinical measurement instruments, rating scales, and so on. In their statistic, the authors suggested replacing the Pearson (or more appropriately, Spearman) correlations in the off-diagonal elements of $\hat{\mathbf{R}}$ with polychoric correlation estimates (or tetrachoric, if the response variables are dichotomous; Olsson, 1979) when computing reliability with Equation 4. The justification for this substitution was that traditional reliability estimates, which rely on Spearman correlations for the categorical response data, underestimate the true population reliability because observed correlations between categorical variables are typically lower than the correlations between the underlying continuous latent variables from which the categorical variables were manifested (see Flora & Curran, 2004, for details). As well, the authors argue that, because Pearson and Spearman correlations (allegedly) require continuous data, the correlation estimates should be replaced by estimates that are not based solely on observed categorical data relationships.

In a small simulation study, Zumbo et al. (2007) demonstrated that, after generating continuous response data using Equation 1 and applying a categorization transformation to create suitable response data (see their associated appendix for the exact coefficients), ordinal α was better at reproducing the population α implied by the continuous underlying variables before the categorization transformation than coefficient α. Based on these results, the authors concluded that ordinal α is more optimal than coefficient α when the response data are categorical, and recommended using ordinal α instead of coefficient α in practice whenever categorical response data are collected. Similar results and interpretations of ordinal α were discussed by Oliden

and Zumbo (2008) and Gadermann et al. (2012), and walk-through material using the R environment (R Core Team, 2016), as well as Mplus (Muthén & Muthén, 2008) and LISREL (Jörsekog & Sörbom, 2006), were presented for applied researchers to use in their analyses.

Misconceptions About Ordinal α

This section presents important misconceptions about ordinal α, as well as coefficient α, borne in the work by Zumbo et al. (2007), Oliden and Zumbo (2008), and Gadermann et al. (2012). These misconceptions are broken into four distinct areas: (1) data requirements to compute coefficient α, (2) the validity of utilizing ordinal α as a measure of reliability with ancillary CTT formulae, (3) the claim that ordinal α provides a better measure of reliability than coefficient α, and (4) the understanding of how ordinal α is situated within modern latent variable theory for psychological measurement.

Misconception 1: Coefficient α Requires Continuous Item Response Data

This misunderstanding likely has multiple sources,[1] but the most notable source relevant to this article can be found in Zumbo et al. (2007). In their article, the authors state that ''coefficient alpha (and KR-20) are correlation-based statistics and hence assume continuous data'' (p. 27). This statement appears again in Gadermann et al. (2012) in multiple locations throughout their article, and also appears in the English abstract in Oliden and Zumbo (2008).

As is clear from the expressions in the previous section, however, all components of CTT, reliability, and coefficient α make no assumptions regarding the distribution or required form of X, X_i, T, or E. All that is required to derive coefficient α, as well as many other concepts related to CTT, is an understanding of covariance algebra with observed variables. Stated differently, any of the aforementioned variables may be discrete or continuous, and can take on any distributional shape within each item and at the composite level. Hence, dichotomous and polytomous item response data are typically valid for computing coefficient α to estimate the test's reliability.

Inspecting this statement again, but this time at a more superficial level, illustrates that Zumbo et al.'s (2007) claim appears to have little coherent meaning. For example, computing Pearson's or Spearman's correlation for dichotomous data results in the ϕ coefficient (Agresti, 2002). Computing the unstandardized correlation (i.e., covariance) between two dichotomous coefficients is no different (see, e.g., McDonald, 1999). Moreover, computing Pearson's correlation between a dichotomous and continuous (or, rather, an interval or ratio scaled) variable is also frequently obtained in practice and is commonly known as the point–biserial correlation (Crocker & Algina, 1986; McDonald, 1999). Most peculiar in the authors' statement above though is the fact that they were obviously aware of the KR-20 formula (Kuder & Richardson, 1937)—a classical reliability statistic that is explicitly intended for dichotomous

response data—and yet still claimed that the observed response data must be continuous at the item level.

To avoid any confusion, and to directly reply to the above claim, we state the following with strong emphasis: *Coefficient α, as well as the KR-20 as a special case, has never required continuous item-level data*. These reliability estimates only require that the observed bivariate relationships among each test item have linear functional forms, and that the observations are coded in interval (or possibly ratio) formats (Stevens, 1946). For dichotomous variables, both of these requirements are true by construction, regardless of the coding scheme.

Additionally, interval data do not inherently require an infinite number of subdivisions in the measured variables (i.e., do not need to be coded with decimal places or fractions). This measurement scale only requires that the distances between commensurate values represent the same quantity (e.g., the difference between 10°C and 20°C is the same as the difference between 70°C and 80°C). This property may have been where the confusions regarding the data requirements for coefficient α occurred in that Zumbo et al. (2007) imply that, because the coding of ordinal variables do not include decimal or fraction values, Pearson or Spearman covariance estimates are invalid. Unfortunately, this is simply not correct, and casts much doubt into the justification for using ordinal α in place of coefficient α for categorical response data.

Covariance Statistics for Ordinal Data. Interest for ordinal α appears in situations where there are three or more ordinal response categories, but the correlation between the ordinal variables may not be optimally captured by a Pearson- or Spearman-based covariance estimate due to the restricted ranges (Olsson, 1979; Rhemtulla, Brousseau-Laid, & Savalei, 2012). Specifically, the observed correlation between ordinal variables (coded as equally spaced intervals) is often slightly lower than the correlation between commensurate interval or ratio variables with a larger number of unique values mainly because Pearson covariance estimates are highly influenced by observations in the tails of the distributions (Fox, 2008). In categorical response data, the range of the observed variables are limited by definition, thus creating difficulty in obtaining higher Pearson correlation estimates. This is also the reason why Spearman's correlation is often lower in magnitude than the Pearson correlation. Note that the relationship between ordered item response data coded as intervals and rank-ordered data is also obvious here, because obtaining Pearson's correlation for ordinal response data that have been rank ordered is equivalent to computing Spearman's correlation for these types of response data.

Rather than assuming that the ordinal variables are on an interval scale, so that a Spearman correlation estimate can be computed, alternative descriptive statistics could be obtained. For example, correlations based on rescaling Pearson's χ^2 test of independence from a two-way contingency table is one suitable correlation alternative. One common statistic for this purpose is known as Cramer's V (Agresti, 2002; Cramér, 1946). Cramer's V is particularly interesting in this respect because when both observed variables are dichotomous the correlation is equivalent to the ϕ

correlation; hence, its use when computing coefficient α from Equation 3 would be equivalent to the classical KR-20 formula. Therefore, using Cramer's V may be a better representation of the observed correlation between two ordinal variables than Pearson's or Spearman's correlation, particularly when the number of response categories is small, and is still in reference to the observed data characteristics.

Unfortunately, however, Zumbo et al. (2007) chose to use an *unobserved* bivariate relationship as a basis for calculating reliability for ordinal variables (i.e., polychoric and polyserial correlations) rather than statistics based on *observed* covariation (e.g., Pearson correlations, Spearman correlations, ϕ, Cramer's V, etc.). Therefore, these authors do not base their reliability estimates on the overt relationship among the test items to describe the composite score X, but rather on the unobserved relationship between the test items for an unobserved composite variable that is unavailable to the test analyst. This has important theoretical and practical consequences, which are further discussed in the three remaining misconceptions and subsequent exploration section.

Misconception 2: Ordinal α Is Equivalent to Other CTT Reliability Estimates

Gadermann et al. (2012) stated the following in their introduction explaining what ordinal α represents:

> Ordinal alpha is conceptually equivalent to Cronbach's alpha. The critical difference between the two is that ordinal alpha is based on the polychoric correlation matrix, described in detail below, rather than the Pearson covariance matrix, and thus more accurately estimates alpha for measurements involving ordinal data. (p. 2)

The above quotation is incorrect on multiple levels; therefore, these statements are broken into Misconception 2 and 3, respectively. Beginning with the first misconception, the conceptual equivalence of ordinal α with coefficient α is obviously incorrect. Simply substituting polychoric correlations into the required matrix to compute coefficient α with Equation 4 fundamentally distorts the meaning of what test reliability is being measured. This is because the supplied correlations are no longer about the *observed* data, but rather the relationship between two *unobserved* continuous variables (typically assumed to have a bivariate normal distribution). Therefore, any important summary statistics, such as the variability of the total score σ_X, cannot be used in concert with any ordinal α estimate because they are in reference to different sets of variables.

To demonstrate, if ordinal α were to be used in the computation of the standard error of measurement (Lord & Novick, 1968) formula

$$\sigma_E = \sigma_X \sqrt{1 - r(X)},$$

where $r(X)$ is replaced with ordinal α, then σ_E may be considerably smaller than it should be for the overt X scores. Using ordinal α would provide unacceptably liberal

sampling error estimates because the reliability estimate would be too high given the observed information provided by X, as well as generate liberal confidence intervals for the test taker's true score given their observed X_i. This will be particularly severe when all of the item response formats contain dichotomous response options, in which case the required tetrachoric correlation coefficients may be notably larger than the related ϕ coefficients (Gadermann et al., 2012). Ostensibly, ordinal α is not conceptually or practically equivalent to coefficient α, and cannot be used as though it were a standard estimate of reliability.

Misconception 3: Ordinal α Provides a Better Estimate of the Population Reliability Than Coefficient α

This misconception appears in both Zumbo et al. (2007) and Gadermann et al. (2012), and generally forms the basis for recommending the use of ordinal α in practice due to its performance in their Monte Carlo simulation. Both sets of authors make this claim within their respective abstract summaries, where Zumbo et al. (2007) concluded that, based on their simulations, ''coefficient alpha is in general a negatively biased estimate of reliability'' (p. 1), while based on the same simulation results Gadermann et al. conclude that ''ordinal alpha more accurately estimates reliability than Cronbach's alpha when data come from items with few response options'' (p. 1). Note that Gadermann et al.'s statement is directly related to the second half of their quotation regarding polychoric correlations in the previous section.

Importantly, however, in their simulation study, Zumbo et al. (2007) either ignored or were simply not aware of a fundamental concept regarding the effects of applying data transformations; namely, that statistical properties prior to the transformation (e.g., variability, covariances, reliability) are not guaranteed (or in most cases even expected) to be invariant posttransformation. As we shall see momentarily, the summary statements in the previous paragraph are in fact invalid because the population reliability coefficient which ordinal and coefficient α were compared to was not with regard to the posttransformed data, but rather to the pretransformed data, and therefore is not a reflection of the test's overt measurement properties. While it is true that coefficient α is a limited estimate of reliability, in that it is understood to be only a lower-bound estimate (Sijtsma, 2009), its limitations are not inherently related to the issue of transforming continuous response data into discrete categories.

In their simulation study, Zumbo et al.'s (2007) final data-generation step was to introduce a data transformation in order to construct the required categorical responses under investigation. However, applying data transformations to the continuous X distribution implied by the relationship $X = T + E$ will necessarily change the distribution of the true scores and errors. Let $f(X) = X^*$ represent the transformation function that discretizes the continuous variable X. For example, to construct dichotomous data one could apply the function

$$f(X_i) = \begin{cases} 1 & \text{if } X_i > \tau, \\ 0 & \text{otherwise,} \end{cases}$$

where τ is some predefined cutoff value. This is one of the transformation functions utilized in Zumbo et al.'s (2007) simulation study, and is easily generalized to polytomous item response formats.

Applying $f(\cdot)$ to both sides of Equation 1 implies that $f(X) = f(T + E) \rightarrow X^* = T^* + E^*$, where the variables T^* and E^* represent the respective transformed variables that equal $f(X)$. Unfortunately, however, because $T^* \neq f(T)$ and $E^* \neq f(E)$ in general, due to the additive relationship between the true score and error variables, the required transformation functions for T^* and E^* are typically impossible to obtain.[2] This property also makes it very difficult to determine what the new population α is after applying the data transformation to X. Moreover, there is no guarantee that the select transformation applied will retain the assumption that $\mathrm{COV}(T^*, E^*) = 0$, which is precisely what occurs in categorical response data due to the dependence of the model-implied expected values and their variances (Lord & Novick, 1968). For example, the sampling variability of a Bernoulli distribution for dichotomous data, relevant to the transformation above, is contingent upon the expected value p_i, where the variance $p_i(1 - p_i)$ is entirely determined by the expected value; hence, the mean and variance are not independent. The same relationship does not necessarily occur in the canonical continuous variable case because the mean and variance can be constructed to be independent.

Based on the above reasoning, it is clear that Zumbo et al. (2007) inappropriately defined what the reliability is at the population level. These authors used the population reliability found in Equation 2 using the continuous data prior to applying the data transformation, where the authors should have used the equation

$$r(f(X)) = r(X^*) = \mathrm{VAR}(T^*)/\mathrm{VAR}(X^*) \tag{5}$$

due to the transformation function $f(\cdot)$. Again, this particular reliability value is often impossible to obtain, even in Monte Carlo simulations, and is only applicable when $\mathrm{COV}(T^*, E^*) = 0$. This is unfortunate because while Zumbo et al. (2007) suggest that the ''measurement model used in [their] simulation involves all of the assumptions of coefficient alpha'' (p. 27) this claim is obviously incorrect after considering the effects of the data transformations. Hence, Zumbo et al. (2007) and Gadermann et al. (2012) were focusing on the wrong population reliability in their simulation study, and therefore many of their conclusions about the performance of ordinal and coefficient α are not only erroneous but also misleading.

One potential saving feature of Zumbo et al.'s (2007) simulation study is that ordinal α is not measuring a test's reliability per se, but rather is measuring *theoretical reliability*—a term that the authors created specifically for their article. Presumably, this term was created to distinguish between a test's observed reliability and its reliability prior to the data transformation required to construct the categorical response data. In this case, ordinal α does recover the so-called theoretical reliability well, as

should be expected given the data generation and estimation approach, while it estimates the observed reliability of the test poorly when the number of response categories is small (see Liu, Wu, & Zumbo, 2010, for similar conclusions with respect to recovering theoretical instead of observed reliability).

Regardless of these result though, the justification as to *why* tests analysts and applied researchers should prefer theoretical reliability over observed reliability has never been discussed. Currently, the author cannot think of any reasonable rationale for preferring theoretical reliability over observed reliability when expressing a test's observed measurement properties.

Misconception 4: Ordinal α Is Supported by Modern Latent Variable Theory

In their discussion section, Gadermann et al. (2012) allude to the fact that ordinal α ''is in line with general current thinking in the psychometric literature about using polychoric correlations for ordinal data'' (p. 7). While this thinking about ordinal data is mainly true from a model-fitting perspective with the normal ogive (i.e., probit) or logistic regression models typically found in item response theory (e.g., Chalmers, 2012; Embretson & Reise, 2000), the computation of a reliability coefficient with a polychoric correlation matrix is not in line with this research field in general.

Despite the removal of statistical information borne from categorizing item response data, ordinal α implicitly assumes that dichotomous test items provide the same amount of statistical information as similar test items that use polytomous or continuous item response formats. This can be seen from the fact that the correlation estimates between dichotomous and polytomously scored items are approximately of the same magnitude as the untransformed continuous variables from which the variables were constructed. On first inspection this property is peculiar because it is well known that truncating continuous variables frequently results in attenuated bivariate correlations due to the loss of statistical information (MacCallum, Zhang, Preacher, & Rucker, 2002). CTT reliability statistics reflect this loss of information in that the reliability estimates for tests constructed from dichotomous items, for example, will systematically be lower than a commensurate test scored using polytomous or even continuous response formats. Ordinal α, on the other hand, provides approximately the same reliability estimate *regardless* of the item response stimuli, generally indicating that the item's method of data collection is of little to no consequence when computing a test's reliability.

With respect to modern statistical measurement theory, ordinal α's implicit assumption that dichotomous, polytomous, and continuous response data provide equivalent forms of measurement information is fundamentally at odds with the concept of statistical information found in more rigorous model-based methods (Embretson & Reise, 2000; Lord, 1980; Samejima, 1969). In item response theory, for example, it is well known that a test item with multiple response options will have more information than dichotomous models with equivalent slope coefficients, and therefore will provide more accurate composite score estimates with smaller

sampling variability (hence, result in more reliable and accurate tests; Baker & Kim, 2004). Clearly, this property is not shared by ordinal α, which currently suggests paradoxical properties of the observed reliability estimates. Note that the paradox generated by ordinal α can, however, be resolved when realizing that the statistic is actually drawing inferences about a *hypothetical* reality rather than the actual reliability of the test (more on this in the next section).

Finally, it is noteworthy to mention that Gadermann et al. (2012) inappropriately cite Green and Yang (2009) for their modern use of the polychoric matrix in computing their reliability estimates. Gadermann et al. (2012) imply that, because this correlation matrix is used by other authors studying reliability, this gives auxiliary support for using ordinal α in practice. However, after inspecting Green and Yang's (2009) article it is clear that these authors explicitly avoid using a reliability definition based on theoretical scores, primarily because this reliability estimate is not useful to the test analyst, and instead resort to an alternative definition based on parallel-forms, which does not suffer from the same conceptual issues. Hence, Green and Yang in no way advertise the use of theoretical reliability estimates such as ordinal α. Gadermann et al. (2012) also cite Bentler (2009) in this regard, though again no formal justification or support can be found for using ordinal α or any reliability estimate based on latent variable scores. In his article, Bentler only stated in passing that drawing inferences about a theoretical reliability estimate was *possible* using his equations, not that this practice is at all recommended in empirical measurement applications that require reliability estimates.

What Is Ordinal α and When Should It Be Used?

Hitherto, the conceptual definition of what ordinal α represents has been lacking due in part to the four aforementioned misconceptions. In this section, we explain what ordinal α truly represents for test analysts, and present an application where ordinal α may be of (limited) interest.

Ordinal α as a Type of Spearman–Brown Prophesy Estimate

The key idea for determining the usefulness of ordinal α lies in understanding the implications of replacing bivariate Pearson and Spearman correlation estimates with polychoric and polyserial correlations. Specifically, replacing the correlation estimates with polychoric estimates results in an $\mathbf{R}$ matrix (call it $\mathbf{R}^*$) that looks and behaves as though all the data were obtained from continuous, bivariate normally distributed data. Indeed, this is the intended purpose of ordinal α, because the covariation between two continuous variables will typically be as high or higher than a Pearson or Spearman correlation between ordinal categorical variables.[3]

Utilizing $\mathbf{R}^*$ to compute coefficient α highlights at least one potentially interesting phenomenon with which ordinal α may be useful. Specifically, ordinal α represents

an *estimate of the expected reliability in an alternative reality whereby categorical responses have been replaced by continuous responses*. This is interesting because it relates to the idea that more than, say, 7 Likert-type response options are not needed for factor analysis applications because the bivariate correlations in **R** (and subsequently, coefficient α) change very little when replaced with polychoric or polyserial estimates (Rhemtulla et al., 2012). In this application, for instance, the Pearson correlation between two 7-point Likert-scale items is likely very similar to the polychoric estimate, and therefore little would be gained by adding more categories in future applications (Flora, LaBrish, & Chalmers, 2012). Hence, the common consensus that more than 7 response categories can generally be considered as continuous data is related to this thought experiment, whereby Pearson correlations may be applied without much loss of generality because the correlation estimates are not meaningfully attenuated (Rhemtulla et al., 2012).

The aforementioned thought experiment also raises a potentially interesting application with respect to ordinal α. In particular, a test analyst may ask: ''If I were to replace [all/some/one] of the categorical items with a continuous response format, how much would I expect coefficient α to increase in future applications?'' This property highlights whether it is fruitful to extend the number of Likert-type response options due to the magnitude of the expected increase from coefficient α to ordinal α. Hence, the question asks about a hypothetical reality where the data collected were ideally continuous for [all/most/some] of the items, and what the expected reliability would be if the categorical item responses were changed to continuous responses in this reality. In essence, this is a type of Spearman–Brown prophecy prediction (Crocker & Algina, 1986), but relates to infinitely increasing the number of response categories within select test items.

As noted by an anonymous reviewer, the above application of ordinal α is only a theoretical approximation. In real-world applications, the skewness of the categorical responses plays an important role in computing the polychoric correlation. For example, an item with three categories and a highly skewed observed distribution could have a lower ordinal α estimate than an item with two categories that has a symmetric distribution (Flora & Curran, 2004). Increasing the number of response categories will only increase coefficient α insomuch as the added response categories do not distort the original categorical distribution of the item response data. Hence, even within this hypothetical Spearman–Brown based thought experiment, the limited usefulness of ordinal α is apparent.

Types of Items Where Ordinal α May Be Justified

The question we are faced with now is whether this hypothetical reality exercise is justifiably applicable to the response data at hand. For some items, such as rating scale or Likert-type response formats, this methodological thought experiment for ordinal α may be reasonable. Consider the following item and response stimuli:

Do you agree that legalization of marijuana is good for the economy?

1. Strongly Disagree
2. Disagree
3. Neutral
4. Agree
5. Strongly Agree

This particular item could theoretically be modified to include more response options, such as ''Somewhat Agree'', ''Mildly Agree'', ''Possibly Agree'', and so on, thereby increasing the number of possible response options that ordinal α implicitly assumes exist.

However, for other item types, particularly in the field of aptitude testing, this mental exercise becomes more difficult to justify. For example,

What is the square root of 100?

1. 100
2. 10
3. 50
4. Cannot be computed

For this item, which is typically scored as correct-incorrect (1-0), increasing the number of response options to mimic a continuous response variable is less clear, and may be inappropriate given the dichotomous scoring scheme. Therefore, applying ordinal α to an item such as this may be less meaningful or justifiable to test practitioners. Although we could add numerous other distractor options to the item (e.g., 90, 10,000, 42, etc.), ordinal α will not reflect this particular modification because it does not pertain to adding more distractor options.

These two examples showcase when ordinal α could, as well as should not, be applied. Other types of items that differ from these two examples will obviously appear in practice, and it is up to the test analyst to determine whether applying ordinal α is justified for each item in the test through careful inspection of each respective item's stimuli. If the response stimuli follow a natural order, as Likert and rating scale items often do, then applying the hypothetical properties implied by ordinal α may be reasonable. However, if the response stimuli cannot be expanded in a simple ordinal fashion then it is likely that ordinal α is inappropriate and should not be investigated or reported.

Discussion

The purpose of this article was to clarify various misconceptions about ordinal α, with hopes of deterring future erroneous claims about a test's overt reliability in practice. Four misconceptions were discussed, and one potentially useful application for ordinal α was presented. Overall, the usefulness of ordinal α appears to be limited to

determining whether more response options for select items with ordinal response stimuli should be included in future data collection samples. However, even in this special application, the potential usefulness of this statistical estimate appears to be minimal in that it is largely limited as an approximate statistical thought experiment. The general conclusion, therefore, is that ordinal α should not be reported as a measure of a tests reliability, but instead should be understood as a distinct theoretical concept.

The results presented herein should not be overly surprising to the reader familiar with reliability theory and statistics in general, particularly if the reader is familiar with the cost of dichotomizing continuous variables or the assumptions required for computing covariance and correlation estimates. Specifically, it is well known that truncating interval or ratio variables decreases the amount of statistical information (MacCallum et al., 2002). Hence, when quantifying scale reliability, we would expect a test which contains categorical response formats to have less information—and therefore lower reliability—than a test in which the same items were expressed as the interval or ratio variables from which the categorizations occurred. Clearly, ordinal α does not reflect this fundamental statistical property when it is considered as an estimate of the test's reliability, and instead only reflects a hypothetical reality regarding how the test may have behaved if the categorical response options were replaced by continuous response formats.

To highlight the issues discussed herein, and to help avoid future confusion among practitioners, the author recommends renaming the current ''ordinal α'' statistic to something more representative of its purpose, such as ''hypothetical α.'' This name not only better represents the underlying meaning of the statistic, but also highlights that the reliability estimate should be interpreted only as a hypothetical estimate of an alternative reality, whereby a test's ordinal categorical response options have been modified to include an infinite number of ordinal response options. This more appropriate name also emphasizes that the ordinal reliability methodology is largely of little use to the majority of practitioners, and may in fact be highly misleading to the intended audience when reported as a measure of a test's overt reliability.

Acknowledgments

Special thanks to two anonymous reviewers for providing insightful comments that improved the quality of this article.

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) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. See McNeish (2018) for a recent presentation of this erroneous claim.
2. Two special transformation are however easy to determine. The first is trivial, where $f(\cdot)$ is simply the identity transformation, denoted $f_I(\cdot)$, which returns the exact values of X, T, and E. The second is the transformation $X^* = f(X + c)$, where c is some constant. This transformation represents a systematic increase in the true scores, which results in higher observed scores, and can be understood by applying the functions $T^* = f(T + c)$ for the true score component and the identity function for $E^* = f_I(E)$.
3. This is not always true as spurious bivariate correlations can occur in practice; see MacCallum et al. (2002).

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.

Bartholomew, D., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. New York, NY: Wiley.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*(1), 137-143. doi:10.1007/S11336-008-9100-1

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

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, & Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin*, *105*, 317-327.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466-491.

Flora, D. B., LaBrish, C., & Chalmers, R. P. (2012). Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Frontiers in Psychology*, *3*, 55. doi:10.3389/fpsyg.2012.00055

Fox, J. (2008). *Applied regression analysis and generalized linear models*. Thousand Oaks, CA: SAGE.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, *17*(3), 1-13.

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*, 155-167. doi:10.1007/S11336-008-9099-3

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

Jörsekog, K. G., & Sörbom, D. (2006). LISREL 8.8 for Windows [Computer software]. Skokie, IL: Scientific Software International.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151-160.

Liu, Y., Wu, A. D., & Zumbo, B. D. (2010). The impact of outliers on Cronbach's coefficient alpha estimate of reliability: Ordinal/rating scale item responses. *Educational and Psychological Measurement*, *70*(1), 5-21. doi:10.1177/0013164409344548

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

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

MacCallum, R., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19-40.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Lawrence Erlbaum.

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

McNeish, D. (2018). Thanks Coefficient Alpha, We'll Take It From Here. *Psychological Methods*, *23*(3), 412-433.

Muthén, L. K., & Muthén, B. O. (2008). Mplus (Version 5.0) [Computer program]. Los Angeles, CA: Muthén & Muthén.

Oliden, P. E., & Zumbo, B. D. (2008). Coefficients of feasibility for ordinal response scales. *Psicothema*, *20*, 896-901.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443-460.

R Core Team. (2016). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria: R Foundation. Retrieved from https://www.R-project.org/

Revelle, W. (1979). Hierarchical cluster-analysis and the internal structure of tests. *Multivariate Behavioral Research*, *14*(1), 57-74.

Rhemtulla, M., Brousseau-Laid, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*, 354-373.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2). doi:10.18637/jss.v048.i02. Retrieved from http://www.jstatsoft.org/v48/i02

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs*, (17). doi:10.1002/j.2333-8504.1968.tb00153.x

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

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677-680.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, *16*(4), 8-14.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods*, *6*, 21-29.