

A Measurement Is a Choice and Stevens' Scales of Measurement Do Not Help Make It: A Response to Chalmers

Educational and Psychological
Measurement

2019, Vol. 79(6) 1184–1197

© The Author(s) 2019

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0013164419844305

journals.sagepub.com/home/epm



Bruno D. Zumbo¹  and Edward Kroc¹

Abstract

Chalmers recently published a critique of the use of ordinal α proposed in Zumbo et al. as a measure of test reliability in certain research settings. In this response, we take up the task of refuting Chalmers' critique. We identify three broad misconceptions that characterize Chalmers' criticisms: (1) confusing assumptions with consequences of mathematical models, and confusing both with definitions, (2) confusion about the definitions and relevance of Stevens' scales of measurement, and (3) a failure to recognize that a measurement for a true quantity is a choice, not an absolute. On dissection of these misconceptions, we argue that Chalmers' critique of ordinal α is unfounded.

Keywords

classical test theory, reliability, Likert item responses, ordinal alpha, ordinal reliability, ordinal data

Introduction

This note is a response to the recent article by Chalmers (2018) concerning alleged misconceptions around ordinal α and related concepts in classical test theory (CTT). It is distressing that an article that claims to address misconceptions in the literature is so full of misconceptions itself. We unpack the most important of these

¹University of British Columbia, Vancouver, British Columbia, Canada

Corresponding Author:

Bruno D. Zumbo, Department of ECPS, University of British Columbia, Scarfe Building, 2125 Main Mall, Vancouver, British Columbia, Canada V6T 1Z4.

Email: bruno.zumbo@ubc.ca

misconceptions in the following sections. Broadly speaking, the misconceptions in Chalmers (2018) can be grouped into three categories: (1) misunderstandings about what are assumptions and what are consequences of a mathematical model, and how these things differ from definitions, (2) misunderstandings about the definitions and relevance of Stevens' (1946) scales of measurement, and (3) a failure to recognize that a measurement for a true quantity is a *choice*, not an absolute.

Like Chalmers, we will mainly refer to coefficient α reliability throughout; however, our remarks are not limited to coefficient α but apply to other quantifiers of test reliability as well. Lest we conflate a critique of ordinal α with more general criticisms of the α reliability coefficient, we remind the reader that Zumbo, Gadermann, and Zeisser (2007) described in the conclusion of their paper a strategy for using the polychoric correlation that could be applied to any reliability quantifier that can be computed from a correlation matrix, such as an ordinal version of the McDonald's coefficient ω (1970, 1999), Revelle's reliability coefficient β (1979), or ordinal coefficient θ among others.

Assumptions, Consequences, and Definitions

Continuous Random Variables Versus Discretizations of Continuous Random Variables

We agree with Chalmers that the calculation of coefficient α does not "require" continuous item response data from a mathematical perspective, in general. However, it does require continuous data if we want α to capture the actual linear variation in an underlying latent continuous process (if one is assumed), rather than simply a discretized proxy of that process. This is the point that was made in Zumbo et al. (2007) when they stated that, "If this assumption is violated, then the Pearson correlation matrix may be distorted" (pp. 21-22).

Chalmers is correct that both the CTT model and the general notion of reliability do not require continuous variables for direct and coherent interpretations but he is mistaken that coefficient α does not. That is, he conflates the general notion of reliability with its quantifier, in this case coefficient α . The reason for this is that α is a function of a covariance matrix, as Chalmers himself shows in his Equation (3). Thus, if we are studying a latent continuous process, then we ideally want this covariance matrix to capture the continuous structure of the underlying process. In contrast, neither the definition of reliability nor the general CTT model rely on a covariance specification.

When computing coefficient α from a set of observed measurements arising from a discretized version of this process, as is the case when using Likert-type items, the covariance matrix will only capture the structure of that particular discretization of the latent phenomenon. While this is intimately related to the underlying continuous structure, it is not the same. This is in fact the exact motivation for Zumbo et al. (2007) to define their ordinal α as they did: Ordinal α attempts to recover some of the continuous structure of the latent continuous process *before* calculating an

estimate of the covariance structure of the items. In this way, one would expect ordinal α to better reflect the hidden, continuous covariance structure of the items. This maneuver does not necessarily recover the *exact* continuous structure of the latent phenomenon, but Zumbo et al. (2007) argued via simulation that it does tend to capture more of the underlying continuous structure than simply ignoring it, as a naive calculation of coefficient α would.

This attempt to recover the continuous structure of latent response processes is also the motivation for the use of underlying variable approaches such as probit regression models, as well as polychoric correlation (e.g., Jöreskog, 1994; Quiroga, 1992), or hybrid approaches in the categorical variable methodology estimator of Muthén (1984) in structural equation modeling and factor analysis in general. Therefore, Chalmers' allegations of the limited usefulness of such an approach would apply equally to this long tradition of widely used underlying variable approaches and the long list of developments that have emerged from this framework.

Mathematically, of course, coefficient α may be calculated regardless of the distributional form of the test data; but it is most accurate to interpret this quantity as a lower bound on the reliability of a test of a continuous latent trait only when the actual measurement (proxy for the latent trait) assumes a continuous form. Otherwise, it is most directly a lower bound on the reliability of a test of a discretized version of this latent trait. Indeed, this fact was another main motivation for the introduction of ordinal α in Zumbo et al. (2007): ordinal α attempts to directly quantify a more appropriate bound on the reliability of a test of a continuous latent trait when only a discretized version of this trait has been measured by testing.

Information Content of Related Measurements

It is false that “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,” as Chalmers (2018, p. 1065) claims. The justification for this claim seems to be only 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.” But this is exactly what should be expected since all these measurements are quantifying the same thing (i.e., a latent continuous trait) in different ways (see section A Measurement Is a Choice; a True Quantity Is Indifferent). This does not prove that any assumption is being invoked.

The fact that these correlations are similar is simply a consequence of the discretization procedure; all measurements (discretized or not) here are quantifying the same phenomenon. The information content in any particular measurement is free to vary depending on the exact measurement procedure, but naturally the information contents of similar measurements will be similar. They need not be identical, and nowhere do Zumbo et al. (2007) claim otherwise. The fact that ordinal α “provides approximately the same reliability estimate regardless of the item response stimuli”

does not indicate that “the item’s method of data collection is of little to no consequence” (Chalmers, 2018, p. 1065). This fact is simply a consequence of how ordinal α proposes to bound reliability; that is, by treating the observed response (discretized or not) as arising from a continuous latent quantity directly.

The Definition of the Classical Test Theory Model

Next, it must be noted that the CTT model of measurement error does *not* assume that $\mathbb{E}(E)=0$ and $\text{Cov}(T, E)=0$, contra Chalmers (p. 1057). These are *consequences* of the definition of the CTT model. This misspecification is in fact a serious mistake and far too common in the published literature. To see why, recall that the CTT model proposes that

$$X = T + E, \quad \text{where} \quad T = \mathbb{E}(X \mid \sigma(f)),$$

where f is an assignment-to-individual function and $\sigma(f)$ denotes the set of (measurable) events generated by this function (e.g., see Kroc & Zumbo, 2019; Zimmerman, 1975; Zimmerman & Zumbo, 2001). In simpler terms, the definition of the true score under the CTT model assures that each individual receives one and only one true score that remains fixed on any actual or hypothetical reapplications of the measurement process X .

The two properties that Chalmers claims are assumptions of the model are easily seen to be consequences of the model under this correct specification. To wit,

$$\begin{aligned} \mathbb{E}(E \mid \sigma(f)) &= \mathbb{E}(X \mid \sigma(f)) - \mathbb{E}(T \mid \sigma(f)) \\ &= T - \mathbb{E}(T \mid \sigma(f)) \\ &= T - T = 0. \end{aligned}$$

Applying double expectation, it follows that $\mathbb{E}(E)=0$. From this, it is easy to calculate

$$\begin{aligned} \text{Cov}(T, E) &= \mathbb{E}(TE) - \mathbb{E}(T)\mathbb{E}(E) \\ &= \mathbb{E}(TE) \\ &= \mathbb{E}(\mathbb{E}(TE \mid \sigma(f))) \\ &= \mathbb{E}(T \mathbb{E}(E \mid \sigma(f))) = 0. \end{aligned}$$

The measurement error model that Chalmers proposes is actually what is known as an *errors in variables* model, and is common in the econometrics literature (e.g., see Hausman, 2001). This is a much weaker measurement error model than that of CTT and lacks the rich structure induced by the CTT model’s individual-level exchangeability of errors condition. This condition is really the key, novel structure of the CTT model; without it, we would not have the defining property that the expectation of the observed score should equal the true score for *every* individual (see Kroc & Zumbo, 2019, for more discussion).

The Definition of Reliability Versus Its Quantification

Researchers and test users often associate the concept of reliability with terms such as “dependability,” “precision,” “repeatability,” and so on, assuming these things are consistent with the mathematical definition in the CTT. In that context, reliability is defined as the ratio of the true score variance and the total variance, or equivalently as, the squared correlation between observed scores and true scores (Gulliksen, 1950; Lord & Novick, 1968; Novick, 1966). Although reliability has been defined in many different ways in test theory, and for most purposes it is immaterial which algebraic expression is taken as a definition and which expressions are regarded as theorems, one advantage of taking the ratio of true score variance and observed score variance as the definition is that it encompasses all observed scores with nonzero variance, whereas the squared correlation is not defined if true score variance happens to be zero.

More generally, Zimmerman and Zumbo (2001) introduced an operator theory formulation of CTT. They described the measurement process as a collection of linear operators acting on a Hilbert space of true score vectors. In this way, the concepts of true score and error score can be naturally associated with projection operators on this Hilbert space. Once this identification is made, metric concepts of distance, length, angle, and orthogonality have immediate implications for test theory. They went on to show, exploiting their operator formalism, that one can consider reliability as a mathematical object that can be defined as another type of projection.

It is this mathematical object, the conventional CTT reliability, that Zumbo and his colleagues call the “theoretical reliability.” The qualifier ‘theoretical’ is appropriate here because this object emerges from the abstract mathematical structure of the underlying (Hilbert) space, and this object is *not* formally estimated in day-to-day psychometric work. Instead, quantifiers like coefficient α simply *bound* (from below) this theoretical reliability (e.g., see Zumbo, 1999).

Although it is commonplace to see the phrase “estimate the reliability” in the psychometric literature, the term “estimate” is deceptive. From a purely formal perspective, one could say that quantifiers like coefficient α are *consistently biased* estimators of theoretical reliability, but this is not what typical practitioners and psychometricians have in mind when they speak about *estimators*. For this reason, it is clearer to say that one may “measure” or “quantify” reliability, via a series of clever and widely used statistical experiments (e.g., repeatable structured data gathering strategies) which, with the aid of mathematical models of the test data such as parallel forms, tau-equivalence, or essential tau-equivalence, can accurately bound the (theoretical) reliability, for example, by means of an empirical copula approach (Bonanomi, Cantaluppi, Ruscone, & Osmetti, 2015).

Because the mathematical object of (theoretical) reliability is defined as a ratio of two components of variance with respect to a population, a given numerical value of “reliability” can be associated with many different combinations of values of true-score variance and error-score variance. To resolve this, one may choose to bound the error term in this quotient and therefore define a quantifier of the reliability by

the design of a measurement experiment. Thus, there are choices one must make, from (1) choosing the manner in which to bound the error, such as internal consistency of a single test administration, interrater variation, or measurement variation over time, to (2) designing of the experiment to actually measure the quantifier of interest, *the estimand*, to (3) the choice of *estimator*. Different estimators naturally yield different properties of their resultant sample *estimates*. Which of these properties are most desirable depends on the objective of the psychometric analysis.

Confusion about estimators (mathematical expressions), estimates (sample values of an estimator), and estimands (target quantities that estimators are defined to quantify) plague most discussions of reliability and coefficient α , and Chalmers' article is no exception. In particular, Chalmers (p. 1066) mistakenly chastises Gadermann, Guhn, and Zumbo (2012) for claiming that Green and Yang's (2009) estimator (bound) of reliability relies on an appeal to the polychoric correlation matrix. He correctly notes that Green and Yang's definition of reliability does not depend on such a quantity but fails to recognize that their definition of reliability defines an *estimand* not an *estimator*. The actual estimator proposed by Green and Yang (2009, Equation 21) utilizes a transformation of discrete test scores into continuous quantities, and the estimation procedure they then employ to compute *estimates* makes explicit use of the polychoric correlation inherited from this transformation (pp. 160, 164, 166).

Stevens' "Scales of Measurement" and Continuity

The Frivolity of Stevens' Scales of Measurement

It is frustrating that so many quantitative social scientists continue to rely on Stevens' (1946) proposed "scales of measurement" as a coherent way to distinguish and categorize measurements. Consideration of these scales is nearly absent in the mathematical and statistical literature of at least the past 30 years, and with good reason: they do not categorize actual measurements in a statistically useful way. Nevertheless, many quantitative social scientists continue to appeal to these scales to try to justify usage or criticism of all manners of methodological choices, often contributing little more than confusion. It is high time to stop this.

Quantitative social scientists often cite Stevens' scales of measurement as a reason for the appropriateness or not of applying a particular statistical procedure to certain kinds of data. Classically, Stevens proposed that one should only consider count and proportion-based statistics for *nominal* data, additionally allowing rank-based statistics for *ordinal* data, mean-based statistics (including covariances and Pearson correlations) for *interval* data, and making no restrictions at all on *ratio* data. What seems to have been lost in the many decades since Stevens' original proposal is the criterion by which he judged the appropriateness of these statistics for these different conceptual types of data. This criterion was *invariance of the statistic under a particular group structure* (Stevens, 1946, p. 678); for example, Stevens argued that statistics deemed appropriate for nominal data should be invariant under permutations of the arbitrary labels one assigns to the nominal categories. However, this particular

criterion that Stevens proposed is *only one of many* different criteria one could imagine. As has been borne out of the decades since Stevens' original proposal, it is clear that many statistics have enjoyed immensely successful usage in a variety of contexts that would be deemed strictly "inappropriate" according to Stevens' criterion.

The most obvious example of this is the fact that no one seems to have any problem with including nominal or ordinal variables as predictors in a regression model. Consider the following simple model:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad (1)$$

where Y is some continuous response (say, height), X is a binary (nominal) variable indicating sex (coded so that $0 = M$, $1 = F$), and $\varepsilon \sim N(0, \sigma^2)$. Naturally, β_0 then captures the average height of males in the theoretical population, $\mathbb{E}(Y | X=0)$, and $\beta_0 + \beta_1$ captures the average height of females, $\mathbb{E}(Y | X=1)$. The corresponding sample statistics will produce the analogous sample estimates of these quantities. Of course, this is exactly what is produced by the ordinary least squares solutions to Equation (1). To see how, recall that the ordinary least squares solutions are

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}, \quad \beta_0 = \mathbb{E}(Y) - \beta_1 \mathbb{E}(X).$$

When X is binary so that $X \sim \text{Ber}(p)$, these expressions reduce to

$$\beta_1 = \frac{1}{1-p} \mathbb{E}(Y | X=1) - \frac{1}{1-p} \mathbb{E}(Y), \quad \beta_0 = \mathbb{E}(Y) - \frac{p}{1-p} \mathbb{E}(Y | X=1). \quad (2)$$

Applying double expectation, we know that

$$\mathbb{E}(Y) = \mathbb{E}(Y | X=0)(1-p) + \mathbb{E}(Y | X=1)p.$$

Plugging this expression into the equations in Equation (2), we recover the natural interpretations of β_0 and $\beta_0 + \beta_1$ as the average heights within each nominal sex group. But these expressions depend on averages and variances of a nominal X , and covariances (correlations) between X and an interval or ratio Y , something that is strictly forbidden under Stevens' original proposal.

A simpler example of the unhelpfulness of Stevens' scales of measurement is that, for binary random variables, proportions are mathematically equivalent to means. Indeed, suppose $\{X_1, \dots, X_n\}$ is a random sample from a $\text{Ber}(p)$ random variable. Then the number of observed "successes" is $\sum_{i=1}^n X_i$, which is equivalent to $n\bar{X}$. Thus, the observed proportion of successes is \bar{X} , the sample mean. Under Stevens' criterion, however, only the first interpretation is allowed, since statistics dependent on the mean are only permissible for interval or ratio data. This clearly illustrates how Stevens' scales of measurement only categorize semantics. They are not mathematically coherent and, in fact, are detrimental to a discussion of statistical usefulness.

A final example of the unhelpfulness of Stevens' scales of measurement is supplied by a statement in Chalmers (2018) that characterizes what he considers to be "Misconception 1" surrounding the use of ordinal α . In his text, Chalmers states the following:

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. (p. 1061)

It is this final line that explains Chalmers' confusion. The confusion is understandable though given that this (incorrect) characterization of interval data has become the industry norm in quantitative social science.

Granting this interpretation for the moment, Chalmers is primarily concerned with binary variables arising within the context of dichotomous item responses on a test. Thus, these variables encode whether a respondent answers each item on the test correctly or not. If we consider a single-item test where $X = 1$ corresponds to a correct answer and $X = 0$ corresponds to an incorrect answer, then X construed as an indicator of correctness is nominal; that is, the proposed encoding is arbitrary and so the structure of X does not change if we simply permute the "correct" and "incorrect" labels. However, if X is construed as the total score of the test, then X appears to be interval according to Chalmers' interpretation since the only "distance" between data points is the distance between "correct" and "incorrect." Trivially then, this distance is commensurate to itself. So, a single dichotomous item test is simultaneously nominal and interval, depending on the semantical interpretation of the information content of the test; that is, does it measure correctness (nominal), or does it measure amount correct (interval)? From a mathematical or statistical perspective this is nonsense, since regardless of the semantics, the information contained in X is the same. Again, we see the frivolity of Stevens' scales of measurement.

Continuous Data

Reexamining the above quote from Chalmers uncovers yet another important flaw in Stevens' scales of measurement. So far, we have blithely accepted Stevens' criterion of *invariance of a statistic under the group transformation appropriate to the scale of measurement*. However, it is not at all obvious what *invariance under a group transformation* meant for Stevens.

Certainly, it did not mean that the value of a statistic would remain unchanged after a group transformation since, for example, the mean, median, and mode will all

change values when scaling a random variable by anything other than unity. It seems what he envisioned was that if one applied the group operation appropriate to the proposed scale of measurement to a set of sample data, then the particular sample data point (realized or hypothetical) that corresponded to the sample statistic would remain the same after the transformation. This is what Stevens seems to imply with the following:

Thus, the case that stands at the median (mid-point) of a distribution maintains its position under all transformations which preserve order (isotonic group), but an item located at the mean remains at the mean only under transformations as restricted as those of the linear group. (p. 678)

From this quotation, it is clear that Stevens' proposed categories of interval and ratio data *can only apply to continuous quantities*. If not, then there is not necessarily any case/item/data point "at the mean." For the dichotomous X considered above, there can never be a sample point at its mean (unless X is essentially a constant) without some kind of linear interpolation, something that Stevens himself noted was not "strictly proper" since "the linearity of an ordinal scale is precisely the property which is open to question" (Stevens, 1946, p. 679). What then would it mean for the sample mean of an ordinal or nominal variable to be invariant under a transformation from the general linear group?

The only apparent way out of this quandary is to recognize that alleged interval or ratio data must arise from only continuous random variables (or certain kinds of discrete-continuous mixtures). Within the context of Chalmers' original criticisms then, requiring interval data to make sense of coefficient α is functionally equivalent to requiring continuous data.

We are hardly the first authors to point out some of the many flaws with Stevens' proposed scales of measurement (see e.g., Mosteller & Tukey, 1977; Velleman & Wilkinson, 1993, or Chrisman, 1998). Yet Stevens' original proposal still clings stubbornly to life in quantitative social science circles. While we recognize that Stevens' work was novel and quite promising in its time, we have learned more than enough in the intervening 70+ years to lay its modern usefulness to rest.

As for the question of determining which statistics are most appropriate for which kind of data, we repeat the advice of Zimmerman (1995) from more than 20 years ago:

Current evidence . . . suggests that the probability distribution of a random variable, not the level of measurement, is paramount in determining which statistical test is appropriate. (p. 93)

To this, we generalize that it is the probability distribution of a random variable, not any purported level of measurement, that should determine which statistic is most appropriate. This statement is the generic justification for preferring the use of

ordinal α over coefficient α for Likert-type measurements of a latent continuous phenomenon.

A Measurement Is a Choice; a True Quantity Is Indifferent

Different Measurements Can Quantify the Same Phenomenon

Perhaps the most distressing misconception in Chalmers (2018) is the failure to recognize that we may choose to measure a true quantity encoded in a random variable in many different ways. Chalmers quibbles with the definition of ordinal α because

substituting polychoric correlations into the required matrix to compute coefficient α . . . 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. (p. 1062)

This statement seems to fail to recognize that we use observed data to study unobserved quantities all the time. Indeed, the entire domain of measurement error is concerned with precisely this enterprise, where a random variable of interest cannot be measured directly, so can only be studied by some observable proxy. It is hardly necessary to point out how lucrative this enterprise has been, and there exist vast arrays of resources summarizing the possibilities (e.g., see Gustafson, 2003, or Kroc & Zumbo, 2019).

Chalmers prefers to use a test composed of dichotomous items to bound the reliability of this test as a measure of a discretized version of the latent continuous process. Zumbo et al. (2007) prefer to use the same test to bound the reliability of the test as a measure of the latent continuous process itself. Here, we see that the two proposals want to use the same measurement process (a test of dichotomous items) to study two intimately related random variables: a latent continuous process or one of its possible discretizations. Both of these propositions are perfectly acceptable from a statistical point of view depending on one's research goals. There is no mathematical, statistical, or conceptual problem with studying some underlying latent phenomenon even if one cannot measure precise realizations of that phenomenon directly. This is what measurement error modelling is for.

However, just because we can compute things, does not mean that those quantities are inherently meaningful, or that they accurately capture the phenomenon we are trying to quantify. In the case of coefficient α applied to a test consisting of Likert-type items, the statistic captures only the structure of the Likert items that are presumed to have discretized a continuous latent process. On the other hand, ordinal α attempts to recapture some of the information in the continuous latent process that has been lost via Likert discretization. In this way, both measures can be seen to quantify the same thing: reliability of the test for the latent phenomenon. Zumbo et al. (2007) simply

argue that ordinal α is a better measure in this context since it recaptures some of the continuity in the latent process.

Changing Measurements Does Not Change True Scores

In this same vein of measurement as a choice, we point out that Chalmers' claim that "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 the errors" (p. 1063) is patently false. This is simple algebra: if one changes the value of one unknown in an equation relating three unknowns, then *at least one but not necessarily both* of the two other unknowns must change. In the context of measurement error, applying a data transformation to the measurement X amounts to proposing a new measurement X' for T . The transformation of X will necessarily change the corresponding errors so that we could now propose $X' = T + E'$.

This of course aligns with reality since each measurement process generates its own error process. For example, one could measure a person's height via any one of the three measurements: (1) use a tape measure and record height to the nearest centimeter, (2) use a yard stick and record height to the nearest yard, or (3) if the person is taller than you are, record 6 feet; otherwise, record 3 feet. All three of these measurements quantify the same phenomenon, and all come equipped with their own error processes. Of course, some of these measurements are better than others at capturing the true quantity of interest.

The same holds in the context of studying a latent continuous phenomenon by means of a discretized (e.g., Likert-type) measurement proxy. If T denotes the latent continuous random variable of interest, and X denotes a discrete (e.g., Likert-type) measurement of this phenomenon, then the corresponding error process E must be continuous for $X = T + E$ to hold. Or, if X' denotes a continuous measurement of the same phenomenon T , then the corresponding error process E' generated by the equation $X' = T + E'$ may be discrete or continuous depending on the particular nature of the measurement process X' . The implicit transformation of observed Likert-type measurements X to continuous proxies X' that characterize the logic of ordinal α is simply a transformation of one measurement process into another intimately related one. The underlying true scores are indifferent to such a procedure. It is precisely because of this transformation that ordinal α will often better capture the structure of a latent continuous phenomenon than will only coefficient α when that latent continuous phenomenon is measured via a Likert-scale proxy.

This general algebraic phenomenon has been exploited in the past, notably by Ekström (2009), to show that the statistical information captured by the phi coefficient is equivalent to that captured by the tetrachoric correlation, and that, under mild conditions, the statistical information captured by Spearman's rank correlation is equivalent to that captured by the polychoric correlation. These results reflect the general reality that the way we choose to quantify (i.e., measure) a particular phenomenon will not change the true underlying value of that phenomenon (observer

effects and quantum entanglement aside). The discretization that occurs when measuring a latent continuous trait by a Likert-response to a certain scale simply changes the measurement X , and corresponding error E ; it does not affect the true score T .

Final Thoughts

Chalmers (2018) proposes four misconceptions surrounding the justification for and use of ordinal α . He claims that (1) coefficient α does not require continuous data to accurately capture the reliability of a test as a measure of a latent continuous phenomenon, (2) ordinal α quantifies reliability of a test as a measure of a latent continuous process rather than of a discretized version of this process generated by the Likert measurements, (3) ordinal α does not provide a better estimate of the population reliability than coefficient α in this setting, and (4) ordinal α is inconsistent with modern latent variable theory because it assumes that the information content of a discretized or continuous measurement of a latent continuous process are the same.

We have seen that Claim (1) is incoherent because of its reliance on an argument from Stevens' scales of measurement. Claim (2) is actually correct and reflects neither a misconception nor a problem of any kind (see section Different Measurements Can Quantify the Same Phenomenon). Claim (3) is empirically refuted in Zumbo et al. (2007), while Chalmers' (2018) theoretical justification for it hinges on an algebra mistake. Both Claim (3) and Claim (4) expose a failure to recognize that one can use many different measurements to quantify the same phenomenon.

Finally, it should be noted that Chalmers (2018, pp. 1067-1068) himself concedes that ordinal α may be the most appropriate quantifier of reliability when using Likert-type measurements to study a latent continuous random variable. Oddly, this is exactly what Zumbo et al. (2007) proposed as the research setting in which ordinal α should apply.

Acknowledgments

We acknowledge the support of The University of British Columbia—Paragon Research Agreement. The authors would like to thank Dr. Oscar L. Olvera Astivia for comments and feedback on an earlier version 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.

ORCID iD

Bruno D. Zumbo  <https://orcid.org/0000-0003-2885-5724>

References

- Bonanomi, A., Cantaluppi, G., Ruscone, M. N., & Osmetti, S. (2015). A new estimator of Zumbo's ordinal alpha: A copula approach. *Quality & Quantity: International Journal of Methodology*, *49*, 941-953. doi:10.1007/s11135-014-0114-8
- Chalmers, R. P. (2018). On misconceptions and the limited usefulness of ordinal alpha. *Educational and Psychological Measurement*, *78*, 1056-1071. doi:10.1177/0013164417727036
- Chrisman, N. R. (1998). Rethinking levels of measurement for cartography. *Cartography and Geographic Information Systems*, *25*, 231-242. doi:10.1559/152304098782383043
- Ekström, J. (2009). *Contributions to the theory of measures of association for ordinal variables* (Unpublished doctoral dissertation). ACTA Universitatis Upsaliensis, Uppsala, Sweden.
- 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. Retrieved from <https://pareonline.net/pdf/v17n3.pdf>
- 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
- Gulliksen, H. (1950). *Theory of mental tests*. Hoboken, NJ: Wiley.
- Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian Adjustments*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Hausman, J. (2001). Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *Journal of Economic Perspectives*, *15*, 57-67. doi:10.1257/jep.15.4.57
- Jöreskog, K. G. (1994). *Structural equation modeling with ordinal variables using LISREL* (IMS Lecture Notes-Monograph Series 24). Retrieved from https://projecteuclid.org/download/pdf_1/euclid.lnms/1215463803
- Kroc, E., & Zumbo, B. D. (2019). *A transdisciplinary view of measurement error models and the variations of $X = T + E$* . Manuscript submitted for publication.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. The Addison-Wesley series in behavioral science: Quantitative methods. Reading, MA: Addison-Wesley.
- McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical & Statistical Psychology*, *23*, 1-21. doi:10.1111/j.2044-8317.1970.tb00432.x
- McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Lawrence Erlbaum.
- Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Boston, MA: Addison-Wesley.
- Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115-132. doi:10.1007/BF02294210
- Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*, 1-18. doi:10.1016/0022-2496(66)90002-2

- Quiroga, A. (1992). *Studies of the polychoric correlation and other correlation measures for ordinal variables* (Unpublished doctoral dissertation). Uppsala University, Uppsala, Sweden.
- Revelle, W. (1979). Hierarchical cluster-analysis and the internal structure of tests. *Multivariate Behavioral Research, 14*, 57-74. doi:10.1207/s15327906mbr1401_4
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680.
- Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician, 47*, 65-72. doi:10.1080/00031305.1993.10475938
- Zimmerman, D. W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika, 52*, 393-408. doi:10.1007/BF02291765
- Zimmerman, D. W. (1995). Increasing the power of the ANOVA F-test for outlier-prone distributions by modified ranking methods. *Journal of General Psychology, 122*, 83-94. doi:10.1080/00221309.1995.9921224
- Zimmerman, D. W., & Zumbo, B. D. (2001). The geometry of probability, statistics, and test theory. *International Journal of Testing, 1*, 283-303. doi:10.1080/15305058.2001.9669476
- Zumbo, B. D. (1999). *A glance at coefficient alpha with an eye towards robustness studies: Some mathematical notes and a simulation model* (Paper No. ESQBS-99-1). Prince George, British Columbia, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioural Science.
- Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficient alpha and theta for Likert response scales. *Journal of Modern Applied Statistical Methods, 6*, 21-29. Retrieved from <https://digitalcommons.wayne.edu/jmasm/vol6/iss1/4/>