Spring 2015

The effects of ordinal data on coefficient alpha

Kathryn E. Pinder
James Madison University

The Effects of Ordinal Data on Coefficient Alpha

Kathryn Pinder

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Arts

Department of Graduate Psychology

May 2015

Acknowledgements

I would like to start by thanking my advisor, Deborah Bandalos, for all of her support with this project. Debbi, you have been an amazing mentor and I cannot thank you enough for all of your feedback, suggestions, and time spent helping me with this project. I feel so privileged to have had to opportunity to learn from you. This project seemed overwhelming at times, but you always helped me refocus, overcome, and make this project the best it could be.

I would also like to thank my committee members, Dena Pastor and Monica Erbacher for their guidance. You provided excellent suggestions and support throughout this process, and I appreciate your taking the time to learn something new with me. An additional thanks to the entire CARS family for their help in talking through problems with me, and being the wonderfully supportive community that you are.

I would be remiss to leave out Cathryn Richmond, Kim Johnson, Sancho Sequeira, and all of my friends for keeping me afloat. Thank you for picking me up when I was down and thank you for all of your love. You have helped me every step of the way.

An additional thank you goes to my family for the immense support and love. I owe a huge thank you to my parents for their unwavering confidence in me throughout my entire graduate career. I cannot thank you enough for everything you have done for me.

Table of Contents

List of Tables

List of Figures

Abstract

Given coefficient alpha's wide prevalence as a measure of internal reliability, it is important to know the conditions under which it is an appropriate estimate of reliability. The present paper explores alpha's assumption of uncorrelated errors when used with ordinal data. Alpha overestimates true reliability when correlated errors are present. In this paper, I use a simulation study to recreate three mechanisms proposed to create correlated errors in ordinal data. The first mechanism, misclassification error, occurs when there are correlated measurement errors present in the data. The second mechanism, grouping error, occurs when there are not enough categories to represent the construct in question. The final mechanism is transformation error, which occurs when observed data do not match the distribution of true scores. Results indicated that misclassification and transformation error caused correlated errors, but only misclassification error caused correlated errors that were large enough for alpha to overestimate true reliability. Researchers should consider the assumption of correlated errors when reporting and making decisions based on alpha's value alone.

Chapter I: Introduction

Coefficient alpha is a popular estimate of a scale's internal reliability. As such, it is reported in nearly every test manual and with nearly every published scale. According to Sijtsma (2009a), over 7,000 papers have citations for alpha, and many more do not cite a source when using the coefficient. Alpha is discussed and reported in journals ranging from highly technical, such as *Psychometrika*, to applied and substantive, such as *Journal of Applied Psychology* and *Personality and Individual Differences*. Oftentimes, alpha is the only reported reliability estimate, and decisions about the suitability of a scale's reliability are made based on alpha's value alone.

Despite alpha's widespread use and popularity, there are a number of assumptions associated with alpha that are unlikely to be met in practice. These assumptions, discussed in more detail in the literature review, are tau-equivalence and uncorrelated errors. In this paper, I focus on the assumption of uncorrelated errors. Correlated errors occur when two or more items on a test share variance above and beyond the variance they share with other items on the test. That is, some aspect unrelated to the relevant construct is causing subsets of items to covary more with each other than they do with the other items on the test. Previous research (in the form of mathematical proofs as well as simulation work) has shown that when errors are positively correlated, alpha overestimates true reliability (e.g., Gu et al., 2013; Raykov 2001). Given alpha's prevalence, it is important that researchers understand the conditions under which alpha is an accurate estimate of reliability. Otherwise, researchers run the risk of making inappropriate conclusions about the suitability of their scales.

Research has established that correlated errors can occur as a result of the format of the test or item wording (e.g., Green & Yang, 2009a). Additionally, many researchers believe that error correlations may occur due to other item or response scale properties, but do not elaborate on the mechanism(s) by which this occurs (e.g., Lucke, 2005; Shelvin, Miles, Davies, & Walker, 2000). The exception is a paper by Johnson and Creech (1983), who suggest that categorizing data from a continuous underlying construct could result in correlated error terms. More specifically, correlated errors could result from the use of ordered categorical data via three specific mechanisms. The first of these mechanisms is called grouping error, and is thought to occur when there are not enough categories to fully represent the construct in question. The second mechanism is transformation error, hypothesized to occur when the underlying continuous distribution does not match the observed categorical distribution of scores in terms of skew. The final mechanism is misclassification error, which occurs when error score elements cause scores to be classified differently than they would have been if they had been classified by true scores only. Each of these mechanisms can cause observed categorized scores to be more related to one another than the true scores would indicate. I expand on these issues in the literature review.

Positive error correlations should cause alpha to overestimate true reliability, but in a pilot simulation study, I found that even with grouping and transformation error introduced into the categorized scores, alpha actually underestimated true reliability for categorized items. This is likely due to the fact that Pearson Product-Moment (PPM) correlations were used in all calculations, and it is well known that PPM correlations are attenuated when used with ordinal data (Bollen & Barb, 1981). In fact, as the number of

categories decreased and the skew of the categorized data increased (theoretically causing increased grouping and transformation error, respectively), alpha underestimated true reliability to a greater extent, which is exactly the opposite of what Johnson and Creech (1983) would predict. This finding perfectly aligns with what Bollen and Barb (1981) would predict, however. Therefore, it is unclear from my pilot study if Johnson and Creech's (1983) mechanisms did cause correlated errors that were smaller in magnitude than the effects of the correlation attenuation, or if these mechanisms did not cause error terms to be correlated at all.

Given alpha's prevalence and the results of the pilot study, it is important to determine if the mechanisms proposed by Johnson and Creech (1983) actually do cause correlated error terms. If so, alpha may not be appropriate for ordinal data. Therefore, in the present paper, I will seek to answer the following questions: 1) Are the mechanisms suggested by Johnson and Creech (1983) actually causing correlated error terms in categorical data? 2) If so, are the correlated error terms causing alpha to overestimate true reliability, or is alpha underestimated due to the correlation attenuation?

To answer the questions above, I will use a simulation study in which I will generate continuous true scores and apply the three categorization mechanisms proposed by Johnson and Creech (1983). I will calculate the true reliability for each dataset in order to estimate the extent to which reliability estimates are biased. Finally, I will examine the extent to which errors are correlated, and the extent to which error correlations (if present) influence the bias of the estimates.

Given the opposing forces of correlation attenuation and correlated errors, it is difficult to predict alpha's performance with categorical data. Nonetheless, I hypothesize

that coefficient alpha will tend to underestimate true reliability, based on the results of the pilot study. I further predict that grouping error and transformation error will not cause correlated error terms in the data, but that misclassification error will.

In the following chapters, I start with a review of the literature relevant to alpha's assumption of uncorrelated errors. This section starts with an introduction to Classical Test Theory (CTT), and the definition of reliability under CTT. Then, I present an alternative way to conceptualize reliability under a Structural Equation Modeling (SEM) perspective. Next, I turn to different ways to estimate reliability, starting with alpha, and then move to SEM-based reliability estimates. Within the section about alpha, I review the following issues: 1) the assumptions of alpha, 2) previous simulation work dealing with the uncorrelated errors assumption, and 3) the role of categorical data. Within the section concerning SEM-based estimates, I discuss previous work that created SEM estimates that do not have some of the same assumptions of alpha. Finally, I present my methods for addressing my hypotheses in detail.

Chapter II: Literature Review

Classical Test Theory and some Definitions

Much of this paper utilizes the framework of Classical Test Theory (CTT). CTT states that for any person's observed score (*X*) on a test item there are two components: 1) a true score (*T*) that represents the person's actual score on that item and 2) an error component (*E*) (Crocker & Algina, 1986). That is, for person *i* and item *j*:

$$X_{ij} = T_{ij} + E_{ij} \tag{1}$$

This equation holds for an individual item *j* on a test and easily extends to a summed test score. According to CTT, true scores for any person and item or test remain the same across test administrations, but *E* scores are completely random. If the same person were to take the same test comprised of the same items repeatedly, without being able to remember their responses from previous administrations, they would obtain different observed scores due to differences in error scores; however the true score for each item (and, therefore, summed test score) would be the same for every administration. The *E* scores, on the other hand, are random and would differ from administration to administration, but at infinitum would sum to zero. Thus, the average of an infinite number of observed scores for a person on an item or a test equals that person's true score on that item or test. Mathematically, this statement is equivalent to Equation 2:

$$\lim_{m \to \infty} \frac{\sum_{j=1}^{m} X_{ij}}{m} = T_{ij} \tag{2}$$

where *m* is the number of administrations. Unfortunately, it is usually not possible to give a person the same test multiple times without some practice effects altering the results. Even if it were possible to eliminate practice effects, a person would need to take

the test an infinite number of times to obtain an accurate true score, which is obviously not possible (Crocker & Algina, 1986). Therefore, a true score can never be directly measured; however, CTT and the theoretical true score will serve as extremely helpful frameworks for the remainder of this paper.

Because true scores cannot be directly measured, other approaches to identifying true score components and error components in observed scores have been developed that rely on alternate test forms. Alternate test forms can be parallel, tau-equivalent, essentially tau-equivalent, or congeneric. In general, the relationship between two tests *c* and *d* can be described by Equation 3:

$$T_c = a_{cd} + b_{cd}T_d + E_{cd} \tag{3}$$

such that a and b are constants, T is the true score on tests c and d, respectively, and E is an error term. This equation describes the relationship between scores on any two tests or any two items on a single test. The properties of a, b, and E depend on the type of alternate test forms.

Parallelism is the most restrictive and strongest relationship between two tests. CTT defines parallel tests as measuring the same construct in identical units of measurement with the same precision (Raykov, 2001). Mathematically, this is equivalent to Equation 3, with the following restrictions (Gu, Little & Kingston, 2013):

$$a_1 = a_2 = \cdots = a_k = 0 \tag{3.1}$$

$$b_1 = b_2 = \cdots = b_k = 1 \tag{3.2}$$

$$\sigma^2(E_1) = \cdots = \sigma^2(E_k) \tag{3.3}$$

for tests with *k* items. As such, parallel tests have equivalent true scores, true score means and equivalent true score and error variances (and therefore equivalent observed score variances).

A weaker relationship between two tests is described by tau-equivalence. When tests are tau-equivalent, they measure the same construct with the same units of measurement, although possibly on a different scale or with a different degree of precision (Raykov, 1997). Tau-equivalent tests meet assumptions (3.1) and (3.2) but may not meet assumption (3.3) (Gu et al., 2013). Thus tau-equivalent tests have equal observed score means and equal true score variances, but unequal error variances and observed score variances. A slightly weaker condition than tau-equivalence is essential tau-equivalence. Essentially tau-equivalent tests meet assumption (3.2), but not necessarily (3.1) or (3.3) (Gu et al., 2013). In other words, true scores from essentially tau-equivalent tests only differ by a constant, resulting in equal true score variances but unequal observed score means. Finally, two test forms can be congeneric. Although congeneric tests measure the same construct, they do so with different units of measurement and precision (Komaroff, 1997). Congeneric tests do not meet any of the assumptions associated with Equation 3 and have unequal true scores and true, error, and observed score variances.

Alternate test forms can be considered as parallel, tau-equivalent, or congeneric, and items within a test can also be described as having those relationships. For example, a test with parallel items would have items with equivalent true scores, true score means and equivalent true score and error variances (and therefore equivalent observed score variances).

Reliability under CTT

Reliability of a scale is an important concept in psychometrics, and an idea that is inextricably intertwined with Classical Test Theory (CTT). Note that this paper focuses exclusively on internal reliability of a scale, and not on alternate forms of reliability such as test-retest reliability or interrater reliability. A scale's internal reliability describes the extent to which that scale is able to measure the construct in question without the influence of measurement error from the use of different questions. In other words, measures of internal reliability quantify the extent to which the use of different scale items contributes to error (Cortina, 1993). Therefore, a scale with high internal reliability has items that are consistent enough that they are not causing test scores to have a large error component. Conversely, a scale with low internal reliability has items that are different from one another in such a way that scale total scores have a large error component. Put differently, error scores are the part of an observed score that are reflective of random effects; that is, random variance unexplained by true scores. If an item is, for example, measuring an irrelevant construct or has idiosyncratic wording, it will be less related to other items on the test, which would harm the internal reliability of a test. CTT assumes that errors are random and therefore uncorrelated, but it is possible that some aspects of error could be systematic. I will explore this idea further in the "Estimates of Reliability" section, below.

Under CTT, reliability is defined as the ratio between true score variance and total test variance (Equation 4). Reliability can equivalently be defined as the squared correlation between the true and observed scores on a test (Equation 5).

$$\rho_x = \frac{\sigma^2(T)}{\sigma^2(X)} \quad (4)$$

$$\rho_x = (r_{XT})^2 \quad (5)$$

In the equations above, reliability increases as true score variance increases (holding other sources of variation constant). Also, true score variance and error score variance are inversely related to one another; given a constant total test variance, as true score variance increases, error score variance decreases. Thus, true reliability is the proportion of total test variance explained by true score variance.

As noted above, however, these equations are strictly theoretical, because true scores are not observable or directly measurable. A more practical definition of reliability uses the idea of parallel tests. According to Sijtsma (2009b), reliability can be defined by the product-moment correlation between two parallel tests. If two tests are truly parallel, then the true components of the scores would correlate, but the random error components would not. Thus if there is a smaller error component, the correlation between the tests and therefore the reliability of each test would be greater. Parallel tests are based on the idea of true scores as well, so it is difficult to show that tests meet the assumption of parallelism. Thus, none of these definitions are practical for objectively measuring and reporting internal reliability for a scale. There are a number of ways to estimate true reliability, although none do so perfectly. These methods are discussed in the section, "Estimates of Reliability." First, reliability is explored from another perspective.

Reliability under SEM

Reliability can be viewed through a structural equation modeling (SEM) perspective as well. According to Sijtsma (2009b), SEM techniques for assessing

reliability are becoming more common as SEM software has become more readily available for researchers. Just as scores can be described in a CTT framework (Equation 1), a score for person *i* and item *j* can be described using an SEM-based equation as well (Equation 6):

$$X_{ij} = \gamma_j T_{ij} + E_{ij} \quad (6)$$

such that γ_j is a factor loading for a given item, given a single-factor scale. To discuss reliability under an SEM framework, it is important to understand the concepts of dimensionality and factor loadings.

Dimensionality refers to the number of factors a scale has. In this paper, the focus will be on unidimensional scales, although many of the concepts can be extended to multi-dimensional scales as well. In SEM, factor loadings reflect the relationship between an item and the construct it is measuring (called a latent variable or a factor in SEM). Factors are not directly observed, but rather are inferred from a combination of observed variables. For example, imagine a five-item scale that claims to measure construct Γ (which could be any number of psychological constructs, such as depression, self-efficacy, sense of belonging, etc.). All five items share some amount of variance, and SEM posits that it is caused by the presence of Γ. All of the items "load" on Γ to a certain extent; a high factor loading indicates that Γ can explain a large amount of the variance in that item. In fact, if an item has a standardized factor loading of γ_j on a factor, then that factor can explain $(\gamma_j)^2$ of the total variance in that item (Raykov & Marcoulides, 2010). In many ways, a factor is like a true score as it does not contain any measurement error. Therefore, factor loadings are directly related to true reliability under

a unidimensional model; larger factor loadings reflect greater amounts of shared variance, and a smaller amount of error, and true reliability is greater.

Under an SEM framework, reliability can be conceptualized as the extent to which items load on a factor (i.e., how much variance in all of the items can be explained by a single factor) as compared to the total variance in the test. Variance in individual item scores that is unexplained by the factor is called that item's "unique variance," which is directly parallel to error score variances in CTT and is represented by the variance of *E* in Equation 6. Therefore, Equation 7 shows McDonald's (1999) omega, an equation for reliability in the SEM framework:

$$\rho_x = \frac{\left(\sum_{j=1}^{k} \gamma_j\right)^2}{\sigma^2(X)} \tag{7}$$

for a unidimensional test with *k* items. If all loadings are equal, this reliability estimate will equal coefficient alpha. Some models have been created to fit scales that have more than one factor (e.g., Raykov & Shrout, 2002), but they are beyond the scope of the present discussion. For a more comprehensive view of reliability in an SEM framework, see Yang and Green (2011).

Estimates of Reliability

Coefficient Alpha.

Coefficient alpha is an extremely widely used indicator of internal reliability. Kuder and Richardson (1937; as cited by Sijtsma, 2009a) developed a form of alpha, but this form was only appropriate for use with dichotomous data. Guttman (1945; as cited by Sijtsma, 2009a) later developed alpha in its current form as one of six ways to estimate a lower bound for reliability. Coefficient alpha became especially popular after

Cronbach (1954) wrote an extensive paper on the estimate, given by the equation below for a test with *k* items:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{j=1}^{k}\sigma^2(X_j)}{\sigma^2(X)}\right] \quad (8)$$

Although alpha does have a number of advantages, such as being simple to compute, alpha also has a number of drawbacks. Most notably, there is a great deal of confusion about how to interpret alpha. As noted by Cortina (1993), many people misinterpret alpha to be a measure of unidimensionality or first-factor saturation, meaning the extent to which all items on a scale are measuring the same construct. Extant literature refutes this claim, however, as high values of alpha can be obtained with a multidimensional scale (Cortina, 1993; Cronbach, 1951).

Assumptions underlying coefficient alpha. Another major drawback of alpha is that it has assumptions that may rarely be met in practice. The first assumption of alpha is tau-equivalence; measures that are not tau-equivalent yield alpha values lower than the measure's true reliability (Lord, Novick & Birnbaum, 1968, pp. 88). Recall from the previous section that tau-equivalence between two test forms requires that they measure the same construct with the same units of measurement. For a single test to be tau-equivalent, each item must measure the same construct with the same units of measurement. Put in terms of Equation 3 and its associated assumptions, consider tests *c* and *d* to be different items on the same test (instead of alternate test forms). For a test to be considered tau-equivalent, two items must only differ in error variances. In terms of SEM, tau-equivalence means that each item must have the same factor loading. A number of simulation studies have found that alpha underestimates true reliability as the number of non-tau equivalent items on a measure increases and as the degree to which

the items violate tau-equivalency increases (e.g., Komaroff, 1997; Zimmerman, 1993). There has been some success in developing estimates of reliability that do not make this assumption, such as coefficient omega (Equation 7). Yang and Green (2011) review other such estimates.

The second assumption of alpha is that error components for each item are uncorrelated. Violations of this assumption are the focus of this paper. Uncorrelated errors are also a tenet of CTT, but this condition is unlikely to be met in practice. Recall Equation 1, which states that for any item on a test, observed scores are comprised of a true score component and error score component. In theory, it is assumed that the error components across items are random and therefore uncorrelated to one another, but this may not be the case in practice. It could be that some subset of items share variance above and beyond the variance they share with the other items on the test; in this case, these items would have correlated errors. Put differently, these items share variance that cannot be explained by the common factor, and that is therefore not reflected in their factor loadings. Raykov (2001) provides a mathematical proof showing that alpha overestimates true reliability when error terms are positively correlated (and underestimated when error terms are negatively correlated), informally shown presently. Recall Equation 8 for alpha:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k}\sigma^2(X_i)}{\sigma^2(X)}\right] \tag{8}$$

Let

$$\alpha = \frac{k}{k-1}[1 - Q] \tag{9}$$

So that

$$Q = \frac{\sum_{i=1}^{k} \sigma^2(X_i)}{\sigma^2(X)} = \frac{\sum_{i=1}^{k} \sigma^2(T_i + E_i)}{\sigma^2(\sum_{i=1}^{k}(T_i + E_i))} \tag{10}$$

$$= \frac{\sum_{1=i,j}^{k} \sigma^2(T_i) + \sigma^2(E_i)}{\sigma^2(\sum_{i=1}^{k} T_i) + \sigma^2(\sum_{i=1}^{k} E_i) + 2\sum_{1=i\neq j}^{k} cov(T_i, T_j) + 2\sum_{1=i\neq j}^{k} cov(E_i, E_j) + 2\sum_{1=i,j}^{k} cov(T_i, E_j)} \tag{11}$$

Because true scores and error scores are truly unrelated to one another, it is true that

$$cov(T_i, E_j) = 0 \tag{12}$$

Substituting into Equation 11,

$$Q = \frac{\sum_{1=i,j}^{k} \sigma^2(T_i) + \sigma^2(E_i)}{\sigma^2(\sum_{i=1}^{k} T_i) + \sigma^2(\sum_{i=1}^{k} E_i) + 2\sum_{1=i\neq j}^{k} cov(T_i, T_j) + 2\sum_{1=i\neq j}^{k} cov(E_i, E_j)} \tag{13}$$

As can be seen in Equation 1), $\sigma^2(X)$ (the denominator of *Q*) contains four components: 1) $\sigma^2(\sum_{i=1}^{k} T_i)$, the sum of variances of each item's true score component, 2) $\sigma^2(\sum_{i=1}^{k} E_i)$, the sum of the variances of each item's error score component, 3)$\sum_{1=i\neq j}^{k} cov(T_i, T_j)$, sum of the covariance between every pair of true scores, and 4) $\sum_{1=i\neq j}^{k} cov(E_i, E_j)$, the sum of the covariance between every pair of error components. Note that the first and second components are also included in the $\sum_{i=1}^{k} \sigma^2(X_i)$ term of alpha (the numerator of *Q*).

When the assumption of uncorrelated errors is met, error terms do not covary, so the fourth component is equal to zero. Thus, *Q* only decreases to the extent that true scores covary with one another. As is clear in Equation 9, when *Q* decreases, alpha increases. In the case of uncorrelated errors, higher alpha values reflect higher true score covariances (equivalently, higher factor loadings and higher true reliability estimates). When the assumption of uncorrelated errors is not met, however, the fourth term is no longer equal to zero, which causes *Q* to decrease (and alpha to increase) to the extent that

error terms are positively correlated; conversely, if negative error correlations are present, *Q* will increase and alpha will decrease. In this way, higher alpha values are not completely reflective of higher true reliability estimates, but are, to some extent, reflective of correlated error terms. Put another way, alpha interprets error covariance as true score covariance, and becomes falsely inflated when positive error correlations are present. Unfortunately, it is impossible to tell from a single alpha value if correlated errors are present or not.

Furthermore, alpha assumes a linear relationship between items. Although it is not usually discussed as an explicit assumption of alpha, alpha does utilize Pearson product-moment correlations, which do assume linearity. This assumption is explored further in the section "Reliability estimates with categorized data," below.

Previous research on the effects of correlated errors. There has been a good deal of simulation work conducted to more fully examine alpha's performance under conditions of error correlations. In 1993, Zimmerman, Zumbo, and Lalonde set the stage for this field of simulation work by generating data such that the true reliability could be calculated directly and compared to estimates of reliability such as alpha. Zimmerman et al. (1993) created data sets of continuous observed item scores that varied on the following characteristics: 1) number of people (observations) varied from 10 to 80; 2) number of items on the scale was set to either 8 or 10; 3) error correlations occurred among either 3 or 6 items; 4) error correlations were set at 0 (not correlated), .25, or .4; 4) true reliability was set to be either .5, .6, .75, .8, or .9; and 5) the distribution of the observed scores was set to be either normal, uniform, exponential, or mixed-normal. Zimmerman et al. (1993) calculated alpha for data resulting from all possible

combinations of these conditions, iterated this process 2000 times, and compared alpha's value to the true reliability estimate. They found that when errors were correlated, alpha overestimated true reliability. Furthermore, there was greater overestimation in datasets with larger error correlations; when true reliability equaled .8 and errors were correlated at .25, alpha was about .863, whereas when errors were correlated at .4, alpha averaged to .873 over all iterations. Alpha also overestimated true reliability to the extent that a larger number of items had correlated errors. When 3 items had correlated error terms, alpha overestimated true reliability by no more than .05, but when 6 items had correlated error terms, alpha overestimated true reliability by as much as .44.

In 1997, Komaroff extended Zimmerman's 1993 research with a very similar simulation study. In Komaroff's (1997) study, continuous data sets representing observed scores were created that varied on: 1) test length (6, 12 or 18 items), 2) factor loadings (0, .2, .5, .7, or 1) on a single factor, 3) correlated errors (item errors were correlated at .2, .5, .7, or 1), and 4) the number of items with correlated errors (1 through half of the items). Consistent with the findings from Zimmerman et al. (1993), Komaroff (1997) found that alpha overestimated true reliability by as much as .66 in the datasets that had a greater number of correlated errors, and the datasets with more highly correlated errors. Additionally, Komaroff (1997) extended previous research by using SEM to estimate the correlated errors and adjusting alpha accordingly, by subtracting the sum of the correlated errors from the numerator and denominator of alpha. This method was effective in adjusting alpha to some extent, but did not fully correct alpha.

Shelvin, Miles, Davies, and Walker (2000) further supported this research. In their simulation study, Shelvin et al. (2000) created continuous observed score data sets

that varied on: 1) factor loadings; all items loaded either .3, .5 or .7 on a single factor, 2) the extent to which two items had correlated error terms (either a correlation of 0, .1, .2, or .3) and 3) sample size (data sets had 50, 100, 200, or 400 cases). As expected, Shelvin et al. (2000) found that alpha increased under conditions of higher factor loadings and higher error term correlations. Furthermore, they found that the correlated errors had a larger effect on alpha when factor loadings were relatively low.

A very simple simulation conducted by Raykov (2001) also supports the finding that alpha overestimates true reliability under conditions of correlated errors. Raykov (2001) varied the degree to which one pair of errors terms were correlated; the correlations were set to be equal to -.4, -.3, -.2, -.1, 0, .1, .2, .3, or .4. Raykov found that increased error covariance lead to alpha slippage. Specifically, positively correlated errors caused alpha to overestimate true reliability and negative correlations among errors caused alpha to underestimate true reliability.

A recent simulation study by Gu et al. (2013) took a more comprehensive view towards both alpha and an SEM estimate of reliability, which was omega (Equation 7) computed with estimates from a non-linear SEM model. Gu et al. (2013) created datasets that varied on: 1) the number of items violating tau-equivalence (either 3 or 6), 2) the ratio of true to error variance (i.e., true reliability; ranged from .1 to .9), and 3) magnitude of error correlations (ranged from 0 to .4). Gu et al. (2013) found that the SEM estimate gave more accurate estimates of true reliability than alpha; overall, alpha tended to grossly overestimate (by as much as .38 when true reliability was low and error correlations were high), and SEM tended to underestimate true reliability, although only by .09 at most. They also found that as the ratio of true to error variance increased, both

estimators demonstrated less bias, and this effect was more pronounced in alpha (i.e., as true reliability increased, both estimators demonstrated decreased bias). Additionally, and as expected based on Raykov's (2001) proof and simulation work, it was found that as error correlation increased, bias of both estimators increased.

In sum, there has been a good deal of simulation work exploring the effects of correlated error terms on coefficient alpha and on SEM-based estimates of reliability. These studies have consistently found that error correlations have a direct effect on coefficient alpha, with positive error correlations causing alpha to overestimate true reliability. Alpha will overestimate true reliability as the magnitude of the correlations increase or as more items have correlated errors. Furthermore, these effects are more pronounced when true reliability (or factor loadings) is relatively low.

The role of categorical data.

All of the simulation work described above was conducted using continuous data, and error terms were forced to be correlated in the data through simulation methods. This raises the question of how error terms become correlated in real data. Green and Yang (2009a) and Lucke (2005) suggest that items could have correlated error terms due to wording effects. For example, if some subset of items on a test is negatively worded, it may cause test-takers to answer in a systematic way that leads to shared variance between those items only. Another example of a wording effect that would cause correlated error terms is items with shared stems or prompts. Again, a particular prompt may be interpreted by test-takers in a particular way, so items that use that prompt will share more variance with one another than they do with the rest of the items on the test. Green

and Hershberger (2000) also point out that memory effects could cause correlated error terms when items build on one another.

Reliability estimates with categorized data.

These are examples of error correlations being caused by certain aspects of the test items, but other researchers have suggested that correlated error terms could arise just by the nature of the data. Imagine a distribution of continuous scores that represent a group of test-takers' true scores, but due to the style of test, test-takers can only answer from a finite number of ordered categories. This type of data is common in the social sciences, and often comes from a Likert-style response scale. Johnson and Creech (1983) suggest that correlated errors are inherent to this type of categorical data. Specifically, they suggest three mechanisms through which ordinal data can cause correlated error terms. The first mechanism is transformation error, which occurs when the distances between categories are not linear transformations of the underlying variable; that is, distributions from the continuous data and the categorical data do not match. For example, if test-takers' true scores formed a normal distribution, but the distribution of observed categorical responses is uniform, this would cause transformation error. The second mechanism Johnson and Creech (1983) describe is grouping error. Grouping error occurs when there are not enough categories to represent the construct. Although some constructs might truly have only a small number of possible levels (e.g., you are pregnant or not), it is likely that many psychological constructs fall on a continuum (e.g., you are somewhere between totally happy and totally sad). When test-takers are forced to classify themselves into one of a small number of categories, the participants' categorized scores will be more similar than their continuous true scores. The final

mechanism suggested by Johnson and Creech (1983) by which errors become correlated in categorical data is misclassification error. The first two mechanisms treat all error as due to classification. Misclassification error is slightly different from the other two sources of error in that it takes into account a continuous error element. Misclassification error occurs when a test-taker's true score would place them into a certain category, but their continuous "observed" score has an error component large enough to cause the score to be categorized differently. See Figure 1 for a graphical representation of misclassification error.

It is worth noting that there has been some discussion about whether correlated errors should truly be considered and treated as "error." Green and Hershberger (2000) argue that some error terms covary systematically due to the item wording or effects from previous items. Since this sort of error would occur on every administration of the test, Green and Hershberger (2000) argue that it is a sort of "reliable error," that does not count as part of a true score (or true score variance), but is not random error either. Rae (2006) agrees with Green and Hershberger's (2000) stance, and argues that correlated errors should be treated as a second factor under an SEM framework, and therefore treated as true score variance; that is, Rae argues that scales' reliability should not suffer from systematic and reliable correlated error terms. Because in this paper I am concerned with error that arises from categorical data, I will not treat systematic error as reliable. It may be appropriate to think of the correlated error terms as comprising a separate "method" factor, but this factor will still be treated as error, and not true score variance.

It is also worth noting that some authors, such as Green and Yang (2009b), argue that the CTT definitions of reliability (Equations 4 and 5) are not appropriate for

categorical data, because they are based on a linear model. Even a scale that has items with categorical data and factor loadings of 1, and therefore no error variance, will yield CTT reliability estimates less than one and factor loadings less than one when Pearson correlations are used with categorical data in calculations (as they typically are). This is due to the fact that ordered categorical data and continuous data differ in their metric; the correlation between the categorical and original continuous scores will always be attenuated. Put differently, Pearson correlations result in underestimates of the relationship between categorical items. Although correlated errors often cause alpha to overestimate true reliability, it is likely that alpha could underestimate true reliability due to attenuated correlations between items due to the categorical nature of the data. These opposing forces will be important to keep in mind throughout this paper.

Alpha with categorical data. Although alpha is purported to be appropriate for use with categorical data according to Guttman (1945, as cited in Sijtsma, 2009a) the effects of categorical data on coefficient alpha have not been thoroughly examined. There are a few exceptions, the first of which is a paper by Lissitz and Green (1975), which included a simulation of categorical data and examined the effects on alpha. In this paper, continuous normally distributed data representing true scores were generated with factor loadings set at .2, .5, or .8, and then uncorrelated error scores were added to create "observed scores." These "observed scores" were then categorized into uniform distributions with 2, 3, 5, 7, 9, or 14 categories, yielding the categorical observed score data sets that were analyzed. Then, alpha and true reliability (calculated as the squared correlation between the true continuous scores and the categorical scores) were calculated. Note that Lissitz and Green (1975) assumed that the error terms for the

categorical data were uncorrelated because the continuous scores did not have correlated errors, counter to what Johnson and Creech (1983) suggest. It was found that as the number of categories increased and as the factor loadings increased, alpha and true reliability increased as well. Moreover, the largest differences between alpha and true reliability were found under conditions of low factor loadings, with alpha underestimating true reliability. This implies that any correlated errors resulting from the categorization process were of a lesser magnitude than the effects resulting from attenuation of correlations due to the use of categorized data. Alternatively, it is possible that the categorization process did not cause error terms to be correlated at all, and the underestimation of alpha is simply reflective of the correlation attenuation. Interestingly, the number of categories did not have an effect on the difference between alpha and true reliability, as would be predicted by Johnson and Creech (1983) due to grouping error. It is impossible from these results to tease apart the respective influences of correlated errors and attenuated correlations on alpha.

A second study by Bandalos and Enders (1996) also examined the effects of categorical data on coefficient alpha's performance. Specifically, they created a simulation study in which continuous data observed scores were created that varied on: 1) the distribution of the continuous data as normal, uniform, moderately non-normal (skew = 1.75; kurtosis = 3.75), severely non-normal (skew = 2.0; kurtosis = 7.0), or leptokurtic and symmetric (skew = 0; kurtosis = 3.0), 2) the distribution of the categorized data (normally distributed, uniformly distributed, and nonnormally distributed), 3) the extent to which items were correlated (25, .5, or .75), and 4) the number of categories (3, 5, 7, 9, or 11). This design allowed for the shape of the categorized data to be either the same as

or different from the shape of the continuous data. This allowed for a test of the effects of transformation error. Bandalos and Enders (1996) found that alpha was highest when the shape of the categorical data's distributions matched the shape of the underlying continuous distribution. This effect was strongest when inter-item correlations were relatively low. Additionally, the authors found that reliability increased as the number of categories increased, but this effect became smaller as the number of categories increased; that is, the difference in reliability was large when moving from 3 to 5 categories, but was minimal when moving from 7 to 9 categories. Johnson and Creech (1996) would expect that as the agreement between the underlying distribution and the categorical distribution increases, transformation error would occur to a lesser extent and alpha would overestimate true reliability to a lesser extent. The exact opposite was found in the Bandalos and Enders (1996) study, as alpha increased when agreement between the two distributions matched, despite true reliability remaining the same across distributions.

As briefly described in the introduction, I found similar results to those of Lissitz and Green (1975) in two pilot studies. In the pilot studies, I created categorized data sets from normally distributed variables that varied on: 1) the number of categories (2, 3 or 5), 2) factor loadings (.4, .6, or. 8), and 3) the skew of observed data (0, 1.5, or 3). According to Johnson and Creech (1983), these variations should cause grouping and transformation error. In the first study, the data sets had 3 items, and in the second, there were 10 items. In both studies, alpha consistently underestimated true reliability to the extent that there were fewer categories and a greater amount of skew, and the effects of skew and categories were greater with lower factor loadings. Additionally, comparison across the two studies revealed that alpha was more biased (that is, underestimated to a

greater extent) when there were only 3 items as opposed to 10 items. Again, it is impossible to tell from these results if there are correlated errors in the datasets that are outweighed by correlation attenuation, or if the attenuation is the only force affecting alpha.

Other Estimates of Reliability

The most common SEM-based reliability estimate for single-factor scales is McDonald's omega (McDonald, 1995). Originally developed as an alternative to coefficient alpha, omega calculates reliability as the variance of the common factor score over the total score variance (i.e., variance of common scores plus variance of unique scores). The numerator of McDonald's omega (Equation 14) is the squared sum of the standardized factor loadings for each item. As the squared factor loadings represent the variance in each item explained by the common factor, the squared sum of these loadings represents the amount of variance that the common factor explains across all items. The formula for omega is below, and is identical to Equation 7.

$$\omega = \frac{\left(\sum_{j=1}^{k} \gamma_j\right)^2}{\sigma^2(X)} \tag{14}$$

As is the case in CTT, omega assumes uncorrelated error terms; however omega is appropriate under conditions of non-tau-equivalence (Revelle & Zinbarg, 2009). When items are tau-equivalent, omega and alpha are equal. Another SEM-based reliability estimate comes from Komaroff (1997), and is presented in Equation 15:

$$\alpha_A = \frac{k}{k-1}\left[\frac{\sum_{i=1}^{k}\sum_{j\neq i,j=1}^{k}\sigma(x_i, x_j) - \sum_{i=1}^{k}\sum_{j\neq i,j=1}^{k}\sigma(e_i, e_j)}{\sigma^2(X) - \sum_{i=1}^{k}\sum_{j\neq i,j=1}^{k}\sigma(e_i, e_j)}\right] \tag{15}$$

such that α_A is alpha adjusted so as to subtract error covariances from the numerator and denominator. This estimate is very similar to omega, except it is adjusted for error

correlations. It does not, however, adjust for violations of tau-equivalence. Thus tau-equivalence is assumed.

The estimates discussed above are linear estimates of reliability; they are only appropriate with continuous data. Recently, some non-linear estimates appropriate for use with ordinal data have been developed. Notably, Green and Yang (2009b) created a non-linear SEM estimate of reliability that has been found to work well when the correct model is specified (Yang & Green, 2011). This estimate is very similar to coefficient omega, but differs in the estimation method used to acquire the factor loadings; Green and Yang (2009b) use a non-linear SEM estimation method that is appropriate for categorical data. Green and Yang (2009b) argue that Equations 4 and 5 are not appropriate for categorical scores, as they would actually be representative of the reliability of the continuous scores, and not the categorical scores. Additionally, correlations among categorized item scores will always be attenuated if Pearson Product-Moment (PPM) correlations are used (Bollen & Barb, 1981). Moreover, many estimates of reliability, including alpha, rely on PPM correlations, which would yield attenuated estimates of alpha when used with categorical data. Thus, alpha might actually underestimate true reliability when data are ordinal. This is precisely what happened in the pilot study.

Green and Yang's (2009b) estimate of reliability circumvents these problems by employing polychoric correlations and weighted least squares estimation, which yield accurate estimates of variable intercorrelations, and therefore of reliability, for categorical data.

Summary

In this chapter, I have discussed two conditions that influence reliability estimates with ordered categorical data: 1) the presence of correlated error terms, which cause alpha and other reliability estimates to overestimate true reliability, and 2) the use of Pearson correlations, which causes many reliability estimates to underestimate true reliability. A great deal of research has been conducted to examine the effects that correlated errors can have on coefficient alpha. A number of simulation studies have been conducted which support Raykov's (1997) proof that correlated errors cause alpha to overestimate true reliability. Nearly all of these studies, however, used continuous data exclusively, and there are two problems with this approach. First, as hypothesized by Johnson and Creech (1983), it may be the case that mechanisms exist by which the categorization of continuous data results in correlated error terms. Second, due to properties of the PPM correlation, estimates of reliability that use this correlation with categorical data are likely to be attenuated. None of the studies I reviewed have addressed these issues in combination. While it is clear that error correlations in continuous data will cause alpha to overestimate true reliability, the same effect in categorical data is largely unstudied. More specifically, the mechanisms through which errors become correlated in ordered categorical data have not been thoroughly studied empirically, nor have the consequent effects on coefficient alpha.

Note that the problems in estimating reliability with categorical data only occur when these data are treated as continuous. When categorical data are treated appropriately (as is the case with Item Response Theory models and non-linear SEM models), these issues do not arise as error terms are not estimated. It is often the case, however, that categorical data are treated as continuous (Flora & Curran, 2004).

Therefore, it is worthwhile to investigate the effects of the source of errors in categorical data being treated as continuous.

Research Questions

I designed the present paper to answer the following research questions, as mentioned previously. First, do any of the three mechanisms proposed by Johnson and Creech (1983) cause error correlations in observed ordinal data? If grouping error is present, errors correlations should be present and should increase in magnitude when there are fewer categories in the observed data. If transformation error is present, errors correlations should arise and increase in magnitude when the observed data are more severely skewed (given that the true continuous scores are normally distributed), and should not arise when the observed data are normally distributed. If misclassification error occurs, error correlations should arise when continuous correlated error terms, representing correlated sources of systematic error, are introduced to scores prior to categorizing, and the magnitude of these estimated error correlations should increase commensurately with those introduced into the continuous data.

The next research question is: if correlated errors are present, do they cause alpha to overestimate true reliability? Are they of great enough magnitude to counteract the opposing effect of underestimation of the PPM correlations? Results from the pilot study suggest that if correlated errors are present, they do not cause alpha to overestimate true reliability, so presumably the attenuation due to use of PPM correlations has a greater effect.

Chapter III: Methods

Data Generation

A simulation study was used to test the hypotheses described above. Observed score data sets were generated that vary on the number of categories and amount of skew, so as to simulate grouping and transformation error as described by Johnson and Creech (1983). Correlated error elements were also added into scores to simulate misclassification error. Several reliability estimates were calculated for each observed score data set to determine which performs best.

Data was generated in SAS 9.4 IML. A 1000 x 1 vector of normally distributed z-scores was generated, and horizontally concatenated 10 times to create a 1000 x 10 matrix of scores, such that each row has the same score repeated 10 times. This matrix represents the underlying true continuous scores for 1000 test-takers on 10 variables; variables can be thought of as items on a parallel test. These scores were weighted by a factor loading (γ) of either .4, .6, or .8, as is common in previous simulation work (e.g., Gu et al., 2013; Komaroff, 1997; Shelvin et al., 2000); all items within a condition had equal factor loadings. At this point, data generation followed one of two paths, described below.

The first path attempted to create all three types of error described by Johnson and Creech (1983). After true scores were generated and weighted by the factor loading, a different 1000 x 10 matrix of normally distributed random z-scores were created to represent error scores, then weighted by

$$\sqrt{1-\gamma^2} \tag{16}$$

such that this weight is inversely proportionate to the factor loading assigned to the true scores and the variance of the total (true score plus error) continuous scores is 1.0. At

this point, data generation followed one of three conditions, which attempted to explore misclassification error. Recall that Johnson and Creech (1983) define misclassification error as correlated error terms resulting from a continuous error element being large enough to "push" some number of scores into a higher or lower category. In the first condition, errors were left as random and therefore uncorrelated; thus, they should not cause misclassification error to occur. In the second condition, three error scores (i.e., the first three columns of the error score matrix) were made to correlate via the Cholesky method at 0.1 or 0.3. The Cholesky method starts with the desired correlation matrix, and uses matrix decomposition to force that pattern in a matrix of raw scores (Fan, Sivo & Keenan, 2002, pp. 206-208). In the third condition, five error scores were made to correlate, again at either 0.1 or 0.3, via the Cholsesky method. These values of error correlation were selected because these are reasonable error correlation values intuitively, and are common in the literature (e.g., Gu et al., 2013; Raykov, 2001; Shelvin et al., 2000). This correlation can be thought of as resulting from item wording effects or testlet effects that cause test-takers to answer certain items in a systematically different way than they would answer other items on the test. In all three conditions described previously, the resulting error scores were added to the true scores, creating a set of "observed" continuous scores.

These "observed" continuous scores were then categorized in one of 18 ways (6 levels of categories X 3 levels of skew, to create grouping and transformation error, respectively). The scores were categorized into 2 through 7 categories, as these are common in practice and were used by Lissitz and Green (1975). Additionally, scores were categorized such that the resulting distribution of categorical scores was either not

skewed (normally distributed), moderately skewed (skew values around 1.5), or severely skewed (skew values around 3). As there has not been research conducted on transformation error, these values were selected somewhat arbitrarily; however the "extreme skew" value was selected specifically to be outside the range of what would be considered to be a normal distribution. The categorization process was accomplished by establishing thresholds along the distribution of z-scores that will give the categorized data the properties described above. For example, to make a dataset with 3 categories that is normally distributed, the thresholds -.674 and .674 would be used; scores below -.674 would be assigned a categorical score of 0, scores between -.674 and .674 would be assigned a score of 1, and scores over .674 would be assigned a score of 2. Thresholds were determined using the cumulative probability density function for the normal curve.

Recall that grouping error occurs when there are not enough categories to represent the underlying construct; thus, errors should become more correlated when there are fewer categories. Transformation error occurs when the distribution of continuous and categorical scores do not match. Therefore, more severe skew in the categorical data should create greater amounts of transformation error, and (subsequently, according to Johnson and Creech, 1983) increased numbers (and magnitude) of correlated error terms. It should be noted, however, that when all items are categorized, an SEM model would not be able to discriminate between error correlations across all items and true factor loading levels. Thus, the second path of data creation (described below) yields data sets with both categorical and continuous items, so as to be able to examine possible error correlations among the categorized items.

The second data generation path attempted to create grouping and transformation errors, but did not attempt to create misclassification error as described by Johnson and Creech (1983). Once true scores were created, a random and appropriately weighted error component (Equation 16) was added to create continuous "observed scores." Then, 3, or 5, items (columns) were categorized into one of 18 conditions using the same method as described above, whereas the remaining 7 or 5 items were left as continuous observed scores, resulting in data sets with both categorical and continuous items.

This data creation and categorization process resulted in 316 different observed datasets for each set of normally distributed continuous true scores (see Figure 2). This process was iterated 1500 times.

Analyses

Analysis of correlated error terms.

A one-factor CFA model was fit to each data set using maximum likelihood estimation, and the parameter estimates were analyzed to determine the extent to which correlated errors were present. These analyses were conducted in M*plus* version 6 (Muthén & Muthén, 2011). Categorized scores were treated as continuous in order to obtain error estimates and correlations among these errors. In the models that were expected to have correlated errors due to the mechanisms proposed by Johnson and Creech (1983), the model allowed for correlated error terms.

To evaluate misclassification error, the all-categorical data sets were analyzed. Recall that these data sets had error elements that were either uncorrelated, or correlated at 0.1 or 0.3 across 3 or 5 items. In the models with 3 or 5 correlated error elements, the CFA model allowed for correlated errors across those items. The error correlations were

examined across these conditions to determine if adding systematic error as Johnson and Creech (1983) predicted did in fact cause correlated error terms.

To evaluate transformation and grouping errors, the data sets with both categorical and continuous data were analyzed. Specifically, in these data sets, the categorical items had error elements that were allowed to correlate, but did not correlate with the error elements from the continuous items. If transformation error is occurring, then the error correlations should increase with increased skew in the observed score data sets. If grouping error is occurring, then the error correlations should increase with fewer categories.

Evaluation of reliability estimates.

For each of the datasets created using the process described above, a number of reliability estimates were calculated. Specifically, the following were computed: 1) true reliability defined by the squared correlation between the observed scores and the true continuous scores, 2) true reliability as defined by the squared correlation between the observed scores and the true scores categorized the same way as the observed data (but categorized without error), 3) true reliability of the continuous data using Formula 14, and 4) coefficient alpha. This first estimate is the value for true reliability that Lissitz and Green (1975) used in their work. The first value is appropriate because it would equal one if error had no effect on the observed scores, so if reliability estimates deviate from one, it would indicate a lack of reliability. In other ways, however, it is an inappropriate true reliability coefficient because it represents the proportion of variance in the continuous true scores (which are actually what we want to compare scores to) that are explained by the categorical observed scores; however this coefficient will never equal

one, even under a situation where no continuous error element was added. This estimate will always be lower than the third true reliability estimate. The second value is appropriate because it would equal one if error had no effect on the observed scores, so if reliability estimates deviate from one, it would indicate a lack of reliability. This was the estimate I used in my pilot study. The third estimate indicates reliability of the continuous data using coefficient omega (McDonald, 1999). Although this represents the true reliability of the continuous data, it does not capture information about the categorization process. Therefore, all "true reliability" coefficients were calculated and interpreted. but I focus on the first two in analyses.

Because true reliabilities were calculated directly, coefficients' performance was compared directly to each true reliability estimate under each of the conditions of data. Specifically, a series of repeated-measures ANOVAs was conducted to examine the extent to which the independent variables (number of categories, skew, loadings, number of correlated error terms, and magnitude of correlated errors) affect the difference between true reliability and each of the reliability estimates.

Chapter IV: Results

I will discuss the results in two sections; first the results from the datasets that contain only categorized items, then the datasets that have results from both categorical and continuous data. Note that I interpret effect sizes (η^2) in all analyses because the statistics are overpowered, as the sample size is 323,930 in the first group of data and 162,000 in the second group of data. I only consider small effect sizes ($\eta^2 > .1$) and larger as meaningful. In the first dataset, the results from the situation in which all items are categorized, there should have been 324,000 rows in the results files; however, in two conditions (those at in which factor loadings were set to .4, there were 2 categories and severe skew, and error correlations were set at .1 or .3 across 5 items), 5 and 9 iterations (respectively) had errors because M*Plus* was unable to estimate several parameters. It is likely that the model did not converge due to a lack of sufficient variance to estimate both the error correlations and residual variances. These rows of data were deleted, leaving 323,986 records. In the second set of data, the results from the partially categorized datasets, there were 162,000 lines of data, as was expected.

Prior to all analyses, I compared all error correlation estimates to one another across all conditions (within each dataset). I wanted to ensure that examining the average of all estimated error correlations within a replication would be appropriate; if the estimated error correlations show the same pattern across conditions, then analyzing an average would be preferable over analyzing each individual estimate. That is to say, I compared the error correlation estimate between x1 and x2 to the error correlation of x1 and x3, x2 and x3, and every other pairwise comparison between present error

correlations to ensure that all estimates were roughly equal within conditions. Visual inspection revealed that all estimates within a condition were approximately equal (only different at the third decimal place or later). Thus, I only analyzed the mean of estimated error correlations in subsequent analyses regarding the Pearson underestimation. Similarly, I compared the standard errors for the estimated error correlations across all present error correlations, and they were also found not to differ between estimates within a condition.

Additionally, I calculated true reliability as a function of factor loadings using Equation 14. When factor loadings were set at .4, .6, and .8, true reliability equals .656, .849, and .947, respectively.

Data Checks

Prior to analyses, a number of aspects of the data were checked to ensure that it was generated correctly. First, factor loadings were checked. On average, factor loadings aligned with the values to which they were set, although slightly underestimated (most likely due to the Pearson attenuation, which is explored further below). Additionally, estimated error correlations were examined. These also aligned with the population values, although they were slightly underestimated. See Table 1 for means, standard deviations, minimums and maximums of estimated parameters. Skew and kurtosis levels were also examined for a random subset of 32,562 data sets. Generally, the skew value was smaller than the value at which it was intended, but there were still notable and practical differences between levels of skew. See Table 2 for means, and standard deviations, of skew and kurtosis values by number of categories.

All Categorical Data

Misclassification error.

The first research question I addressed was whether misclassification error caused the data to have correlated error terms. Recall that misclassification error occurs when continuous and correlated error elements are introduced into true scores prior to categorization. In the present study, I introduced misclassification error by adding error terms across either 3 or 5 items correlated at either 0.1 or 0.3 to true continuous scores. If misclassification error occurs, as Johnson and Creech (1983) would predict, errors for pairs of categorical items will become more correlated when the correlation between the continuous error elements increases. When 3 items had correlated errors, error correlations did arise. When errors were set to correlate at 0.1, average error correlations were approximately .068 (*SD* = .05). When errors were set to correlate at 0.3, average error correlations were approximately .21 (*SD* = .06). The average of the estimated standard errors for these error correlation estimates was .05 (*SD* = .02), for both levels of error correlation. Correlated errors were also present when 5 items had correlated errors. When error correlations were set at 0.1 across 5 items, the average estimated error correlations were approximately 0.07 (*SD* = .06). The estimated average standard error for these error correlation estimates was .05 (*SD* = .02). When error correlations were set at 0.3 across the 5 items, the average estimated error correlations were approximately 0.21 (*SD* = .07). The estimated average standard error for these error correlation estimates was .02 (*SD* < .001). This pattern was practically the same across the three levels of factor loadings; the interaction between loadings and estimated error correlations had a negligible effect size (Table 3). This finding supports the idea that

misclassification error occurs. Furthermore, and as demonstrated above, average estimated error correlations were the same regardless of the number of items set to have correlated errors.

The estimated error correlations were lower than the values at which they were set in the continuous data, and this is likely a result of the fact that this categorical observed data was analyzed as continuous; that is, the Pearson correlations were underestimated due to the ordinal nature of the data. If this is true, then the estimated Pearson correlations will be closer to the true correlations (that is, the correlations I set) as the number of categories increases and as the observed categorical data is more normal (given that the continuous "observed" scores are normally distributed).

I ran a between-subjects ANOVA to determine the influence of skew and number of categories on estimation of error correlations (Table 4). The two-way interaction between skew and number of categories was significant but the effect size was very small ($\eta^2 = .008$). The main effects of skew and categories were both significant and had effect sizes of $\eta^2 = .004$ and $\eta^2 = .015$, respectively. Generally, as the number of categories increased, the estimate of the error correlation increased (i.e., became closer to the value at which it was set; see Table 5 and Figure 4). Similarly, when less skew was present, error correlation estimations increased. When there were five categories, however, this pattern did not hold, and all estimates of error correlations were roughly equal across levels of skewness. It is possible that this effect is due to the fact that skewness in the 5 category condition was notably different that the 4 and 6 category conditions (Table 2). The pattern of increased estimated error correlations with increased skew was true

regardless of whether error correlations were set at 0.1 or 0.3. These results indicate that underestimated error correlations are likely due to the use of Pearson correlations.

Effects on alpha.

The second research question I addressed was whether these correlated errors caused alpha to overestimate true reliability. Specifically, alpha should overestimate true reliability to a greater extent when more items have correlated errors and when those correlations are greater in (positive) magnitude. I ran a 3 x 2 mixed ANOVA using difference contrasts (comparing alpha to each reliability value) to examine this hypothesis (Table 6). Both estimates of true reliability were always lower than alpha, and the first true reliability estimate, the squared correlation between observed categorical scores with categorized true scores, was always lower than the second reliability estimate, the squared correlation between observed categorical scores and true continuous scores (see Table 7 and Figure 4). Furthermore, both true reliability estimates were greater in magnitude when only three items had correlated errors, but alpha was higher when there were five correlated error terms. Additionally, both true reliability estimates decreased when errors were more strongly correlated, but alpha increased when errors were more highly correlated. Despite the small effect sizes, these results align with the hypothesis that alpha would overestimate true reliability when there were more items with correlated errors and the magnitude of the correlations was greater. Keep in mind that the actual estimated error correlations depended on the level of skew and the number of categories, and Figure 4 is based on the means of estimated error correlations across level of skew and number of categories. To see a graphical demonstration of how skew and categories affect alpha and reliability coefficients, see Figure 5, and see Table 7 for means.

These analyses also revealed that alpha and true reliability estimates significantly differ as a function of loadings (see Table 8 for means). Alpha and both true reliability estimates increase with higher factor loadings.

Partially Categorized Data

Grouping and transformation error.

The first research question is whether the categorization process caused correlated errors. Johnson and Creech (1983) would predict that in the categorized items, error correlations would be present and would increase in magnitude with fewer categories (grouping error) and more skewed observed data (transformation error: assuming normally distributed true score data, which is the case in the present paper).

I ran a three-way between-measures ANOVA to determine if the average estimated error correlations differed as a function of the loadings, number of categories, and skew in the observed data (see Tables 9 and 10). The most meaningful result from this ANOVA was the significant interaction between skew and loadings on the estimated value of error correlations, $\eta^2 = .10$. When factor loadings were low, skew did not have much of an effect on estimated error correlations, but as loadings increased, the effect of skew on estimated error correlations became more pronounced; this finding is explored further in the discussion section. Estimated standard errors for error correlations were .04 for all conditions. See Figure 6 for a graphical representation of these results. These results support the presence of transformation error; but the extent to which transformation error occurs depends on the amount of skew. The results do not support the presence of grouping error because had grouping error been present, error correlations

would have been present and increasing in magnitude as there were fewer categories, and this pattern did not occur.

Effects on alpha.

As correlated errors were present, I examined if the correlated errors would cause alpha to overestimate reliability. To do this, I ran a 2 x 3 mixed ANOVA that compared alpha and the first reliability value as a function of skew, loadings, and categories (Table 12). In this case, I only analyzed the first true reliability estimate because the data were only partially categorized, so it would have been inappropriate to correlate these scores with entirely categorical true scores. .The results of the ANOVA revealed that there was a significant between-measures main effect of factor loadings; all estimates increased as factor loadings increased. Interestingly, despite the fact that correlated errors were present due to skew (as described above), alpha only very slightly overestimated the first true reliability estimate in the conditions of moderate and severe skew. See Figure 7 for a graphical depiction of these results.

Chapter V: Discussion

These results partially support the mechanisms proposed by Johnson and Creech (1983). Specifically, misclassification error and transformation error were found to produce correlated errors, but not grouping error. Moreover, the magnitude of correlated errors affected alpha's value, such that alpha was greater when there were higher correlated errors; however the correlated errors were only large enough to cause alpha to overestimate true reliability in certain conditions. I discuss each of the mechanisms proposed by Johnson and Creech (1983) in turn, as well as their effects on alpha in relation to true reliability.

Misclassification Error

Recall that misclassification error was defined as error correlations resulting from measurement error, which systematically places people in a higher or lower category than their true score would dictate. According to Johnson and Creech (1983), correlated errors should arise in observed categorical data to the extent that systematic error measurement occurs. That is, when systematic measurement error causes relatively higher correlations across a subset of items, the error correlations in the observed data should increase commensurately. The results support the presence of misclassification error. When continuous measurement error was made to correlate at a lower value, the estimated error correlations were relatively low, and when continuous measurement error increased, the estimated error correlations increased as well. Furthermore, alpha was found to overestimate true reliability to the extent that errors were correlated; alpha overestimated true reliability to a greater extent when there were more items with correlated errors and those errors were greater in positive magnitude. This supports much of the previous

research, including that by Zimmerman et al. (1993), Komaroff (1997), and Shelvin et al. (2000). Notably, both true reliability estimates decreased with larger and more error correlations. This makes sense because these values do not conflate error covariance with true score covariance (as alpha does). Recall that alpha assumes error covariances to be zero, and increases to the extent that error correlations are increasingly positive. Thus, misclassification error is something researchers should be aware of when computing alpha with categorical data.

Transformation Error

Johnsons and Creech (1983) define transformation error as error correlations arising due to a nonlinear transformation of continuous scores to observed categorical scores. In the present study, I introduced transformation error by categorizing the normally distributed continuous "observed" score data in a way to force skewness. Correlated errors did arise when observed data were skewed, and they were of greater magnitude when data were more severely skewed. This finding supports the presence of transformation error. When there was no skew present (and therefore no or very small correlated errors) or factor loadings were .4 or .6, alpha was equal to the first true reliability estimate. Recall that the first true reliability estimate is the squared correlation between continuous true scores and observed categorized scores. When skew was present and there were high factor loadings (the condition under which correlated errors were highest), alpha did slightly overestimate the first true reliability estimate. Surprisingly, alpha overestimated the first true reliability estimate to a greater extent when there was moderate skew than when there was severe skew (only in the highest factor loading condition).

There is evidence that the Pearson attenuation problem occurred as well in relation to transformation error. As would be expected, alpha was lower when the data were more severely skewed. This is likely a result of the fact that Pearson correlations underestimate relationships to a greater extent when data are more skewed (Bollen & Barb, 1981). This effect may explain why alpha overestimated the second true reliability estimate to a greater extent when there was moderate skew as opposed to severe skew. It may be that the Pearson attenuation effect had a relatively larger effect on the severely skewed items than on the moderately skewed items. That is, the Pearson attenuation effect balanced out the overestimation due to correlated errors in the condition of high skew to a greater extent than it did in the condition of moderate skew.

There was also one surprising result in regard to transformation error. As shown in Figure 4, the effect of skew on estimated error correlations increases when factor loadings are higher, which replicates a result from Johnson and Creech (1983). This is possibly due to the fact that SEM models imply a linear relationship between factor scores and the probability of responding to an item a certain way. Probabilities should be modeled with an s-shaped curve with asymptotes at 0 and 1 (as is done in Item Response Theory). When factor loadings are relatively higher, the curvature of that s-shaped curve becomes more extreme and looks less linear. Thus, the SEM linear model becomes more inappropriate. The residuals around the line are skewed to the extent that the line is an inappropriate model. Thus, the correlations of the residuals (i.e., the error correlations) increase due to the fact that they have more skewed distributions.

Grouping Error

According to Johnson and Creech (1983), grouping error occurs when correlated errors arise from there being too few categories to represent the construct. They hypothesize that error correlations would increase as the number of categories in the observed categorical data decrease. In the present study, I attempted to introduce grouping error by varying the number of categories from two through seven to see if correlated errors would arise and increase in magnitude with fewer categories present. Grouping error did not cause errors to be correlated, as Johnson and Creech (1983) hypothesized. The number of categories in the observed data set did not systematically cause estimates of correlated errors to be higher when data had fewer categories. In fact, estimated error correlations were generally highest when six categories were present in the data, and lowest when data only had two categories. This is likely a result of attenuation of Pearson correlations, which would dictate that Pearson correlations are smaller when there are fewer categories (Bollen & Barb, 1981). The data follow this pattern more closely than the grouping error pattern that Johnson and Creech (1983) predicted, which dictates the opposite pattern. The effects of the number of categories present in the data were small overall, indicating that the two opposing forces (error covariation and attenuation) may have, for the large part, cancelled one another out.

Limitations and Future Research

There are several limitations in this study. First, the conditions of skew were somewhat arbitrarily decided, and were not very severe (even the condition with the most severe skew has an average skew level of 2.78). Further research should examine the effects of more severe skew, although I expect the same patterns found in the present paper will continue to be present. Additionally, the levels of skew were not consistent

across the number of categories. Although the average skew level across categories was appropriate, some categories had markedly less skew than others within a condition of skew (Table 2).

The study also sacrificed some generalizability in favor of being able to more directly assess the mechanisms proposed by Johnson and Creech (1983). First, the data I generated met some strict assumptions. For example, the generated true score data were made to be parallel, which is a strict assumption, and is also an assumption of alpha. Future research may examine the joint effects of non-tau-equivalence and error correlations. Additionally, like much of the current research on alpha and internal reliability, this study only examined items with a unidimensional structure. Some work has been conducted to determine SEM-based reliability estimates in the case of multidimensional tests (Green & Yang 2009a), but the effects on alpha specifically have not been thoroughly examined. Also, for the sake of identifying grouping error and transformation error, I created data that contained both categorical and continuous items. This type of data is unlikely to occur in practice, as most scales have a consistent response scale. Thus, effects from grouping or transformation error may be hard to disentangle from true score variance in practice.

Relatedly, this study only analyzed coefficient alpha, whereas there are many other reliability estimates. It would be beneficial to examine the performance of other reliability estimates when correlated errors are present; most notably, McDonald's (1999) omega under conditions of non-tau-equivalence, and Gu et al.'s (2013) non-linear omega.

Summary

Taking these limitations into account, there are still several important implications from this research. Most notably, error correlations in the observed data can arise when measurement errors (in the CTT sense) are correlated. Error correlations can also arise when observed data do not match the distribution of true scores. Furthermore, error correlations do cause alpha to overestimate true reliability. When observed data are ordinal, this overestimation can be attenuated or negated by the use of Pearson product-moment correlations. According to the results of the present study, correlated errors caused by misclassification error (i.e., correlated measurement error) are large enough that alpha always overestimates true reliability. When alpha is computed from skewed data, it also has the potential to overestimate true reliability, although this effect is exaggerated with higher factor loadings. Overall, the results suggest that alpha is only appropriate with ordinal data when data are not skewed and do not have systematic measurement error; however it appears to be appropriate with any number of categories. Additionally, the number of categories does not interact with loadings, skew or measurement error; these effects do not become more or less present with different numbers of categories.

Chapter VI: Conclusions

This study addressed some proposed mechanisms by which errors become correlated in ordered categorical data. Specifically, I simulated misclassification error, grouping error, and transformation error, as described by Johnson and Creech (1983). In short, I found support for the occurrence of misclassification and transformation error, but not for grouping error. That is, correlated errors arose when correlated measurement error was present and when observed data did not match the underlying normal distribution; however correlated errors were not strongly affected by the number of categories.

Given the results of the current study, researchers should be especially cautious in using alpha if their data are skewed when a normal underlying continuum is assumed; for example, with a scale measuring psychological constructs such as ability given in a gifted class, such that data are negatively skewed. Researchers should also be especially careful if they have reason to believe that some items may be sharing extra variance, due to properties of the test or items. Using alpha when the assumption of correlated errors is not met and data are ordinal in nature may lead researchers to come to inappropriate conclusions about the reliability of their scale. Based on the current literature (including the results of this study), it is not always the case that alpha is a lower or upper bound for true reliability. Thus, researchers need to take caution when reporting alpha and note possible violations of assumptions.

Coefficient alpha is an extremely popular measure of internal reliability, and is reported in many test manuals and journals. Additionally, many scales and tests yield ordinal data, either in dichotomous or a Likert-type format. Therefore, researchers

should be aware of alpha's potential for inaccuracy in these situations. In particular, alpha is often applied to ordinal data, and such data are subject to the joint effects of attenuated inter-item correlations, resulting in lower values of alpha, as well as correlated errors, resulting in higher values of alpha. It is impossible to tell if alpha is over- or underestimating reliability from a single alpha value. Additionally, it is common belief that a test must show reliability before it can be shown to have validity (Crocker & Algina 1986, p. 217). If researchers base their decisions about a scale's reliability on alpha's value alone, they risk misinterpreting the validity of that scale as well.

As Sijtsma (2009a) mentioned, SEM software and knowledge is becoming more readily available to applied researchers. I recommend that researchers conduct SEM analyses on their data using the appropriate model for their data (i.e., with appropriately defined parameters and appropriate estimation methods for the type of data) to ensure that the assumption of uncorrelated errors is met. If the assumption is not met, I recommend using a different reliability estimate. Although it has not been thoroughly researched or used in practice yet, the non-linear omega developed by Gu et al. (2013) shows promise as an accurate reliability estimate for use with ordered categorical data that assumes an underlying continuum.

Understanding that SEM estimates are not an option for many applied researchers, I (again) recommend that researchers who do report alpha also note possible violations of assumptions, even if they cannot be explicitly tested. As I demonstrated in the current study, and has been demonstrated in previous research, alpha has potential to over- or underestimate true reliability in the presence of correlated measurement error and skew in

observed ordered categorical data. Thus, researchers should always be skeptical of alpha's value as an appropriate estimate of internal reliability.

References

Bandalos, D. L., & Enders, C. K. (1996). The effects of nonnormality and number of response categories on reliability. *Applied Measurement in Education*, *9*(2), 151-160.

Bollen, K. A., & Barb, K. H. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review*, 232-239.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98.

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Fan, X., Sivo, S., & Keenan, S. (2002). *SAS for Monte Carlo studies: A guide for quantitative researchers*. Sas Institute.

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*(4), 466.

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling*, *7*(2), 251-270

Green, S. B., & Yang, Y. (2009a). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *74*(1), 121-135.

Green, S. B., & Yang, Y. (2009b). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*(1), 155-167.

Gu, F., Little, T. D., & Kingston, N. M. (2013). Misestimation of reliability using coefficient alpha and structural equation modeling when assumptions of tau-equivalence and uncorrelated errors are violated. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(1), 30.

Johnson, D. R., & Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, 398-407.

Komaroff, E. (1997). Effect of Simultaneous Violations of Essential τ-Equivalence and Uncorrelated Error on Coefficient α. *Applied Psychological Measurement*, *21*(4), 337-348.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology*, *60*(1), 10.

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

Lucke, J. F. (2005). "Rassling the hog": the influence of correlated item error on internal consistency, classical reliability, and congeneric reliability. *Applied Psychological Measurement*, *29*(2), 106-125.

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

Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

Rae, G. (2006). Correcting Coefficient Alpha for Correlated Errors: Is α K a Lower Bound to Reliability? *Applied Psychological Measurement*, *30*(1), 56-59.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*(2), 173-184.

Raykov, T. (2001). Bias of Coefficient α for Fixed Congeneric Measures with Correlated Errors. *Applied Psychological Measurement*, *25*(1), 69-76.

Raykov, T., & Marcoulides, G. A. (2010). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.

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*(2), 195-212.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145-154.

SAS Institute Inc. 2013. *SAS® 9.4 Guide to Software Updates.* Cary, NC: SAS Institute Inc.

Shevlin, M., Miles, J. N. V., Davies, M. N. O., & Walker, S. (2000). Coefficient alpha: A useful indicator of reliability? *Personality and Individual Differences*, *28*(2), 229-237.

Sijtsma, K. (2009a). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107-120.

Sijtsma, K. (2009b). Reliability beyond theory and into practice. *Psychometrika*, *74*(1), 169-173.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for t he 21st century?. *Journal of Psychoeducational Assessment*, *29*(4), 377-392.

Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement*, *53*(1), 33-49.

Table 1

Estimated Parameter Descriptives

			M	*SD*	minimum	maximum	*n*
Fully categorized data	Loadings	0.4	0.329	0.179	-0.36	0.72	107986
		0.6	0.495	0.183	-0.51	0.76	108000
		0.8	0.706	0.089	-0.64	0.83	108000
	Error correlations	0.1	0.068	0.037	-0.72	0.40	161991
		0.3	0.212	0.051	-0.100	0.48	161995
Partially categorized data	Loadings	0.4	0.344	0.080	-0.3	0.44	80999
		0.6	0.523	0.106	-0.5	0.63	80994
		0.8	0.720	0.057	-0.61	0.81	80947
	Error correlations		0.061	0.081	-0.080	0.370	161965

Table 2

Skew and Kurtosis Values by Number of Categories and Intended Level of Skew

Number of categories			No skew	Moderate skew	Severe skew
2	skew	*M*	0.001	2.134	3.496
		SD	0.037	0.075	0.129
	kurtosis	*M*	-2.000	2.578	10.311
		SD	0.000	0.321	0.918
3	skew	*M*	0.001	2.085	3.549
		SD	0.020	0.059	0.117
	kurtosis	*M*	-0.990	3.504	12.132
		SD	0.024	0.278	0.929
4	skew	*M*	-0.001	1.280	2.089
		SD	0.023	0.043	0.062
	kurtosis	*M*	-1.102	0.321	3.446
		SD	0.016	0.131	0.298
5	skew	*M*	0.001	2.064	3.353
		SD	0.017	0.045	0.093
	kurtosis	*M*	-0.506	4.365	12.296
		SD	0.030	0.259	0.788
6	skew	*M*	0.000	2.077	2.134
		SD	0.028	0.057	0.053
	kurtosis	*M*	-1.400	3.584	4.760
		SD	0.003	0.287	0.304
7	skew	*M*	-0.001	1.482	2.055
		SD	0.031	0.041	0.056
	kurtosis	*M*	-1.598	1.377	3.713
		SD	0.015	0.151	0.284

Table 3

Eta-squared Values for the Effects of Magnitude of Error Correlations and the Number of Items with Correlated Errors on the Estimate of Correlated Errors

	Sum of Squares	η^2
Magnitude of error correlations	1696.216	0.195
Number of items with correlated errors	0.019	0.000
Magnitude*Number of items	0.002	0.000
Error	651.243	
Total	8687.832	

Table 4

Eta-squared Values for the Effects of Number of Categories, Skewness, and Factor Loadings on the Estimate of Correlated Errors

	Sum of Squares	η^2
Number of categories	128.159	0.0148
Skewness	35.473	0.0041
Factor loading	1.733	0.0002
Categories*Skew	16.619	0.0019
Categories*Loading	0.86	0.0001
Skew*Loadings	11.452	0.0013
Categories*Skew*Loadings	1.867	0.0002
Error	2151.262	
Total	8687.832	

Table 5

Average Estimated Error Correlations by Skew and Number of Categories in Entirely Categorical Data Sets

Number of Categories	Skew	*M*	*SD*	*n*
2	none	.118	.067	1800
	moderate	.101	.063	18000
	severe	.086	.064	17986
	average[a]	.102	.066	53946
3	none	.146	.079	18000
	moderate	.126	.074	18000
	severe	.108	.072	18000
	average	.127	.077	54000
4	none	.167	.087	18000
	moderate	.150	.082	18000
	severe	.135	.077	18000
	average	.151	.084	54000
5	none	.173	.090	18000
	moderate	.150	.084	18000
	severe	.130	.082	18000
	average	.151	.087	54000
6	none	.143	.078	18000
	moderate	.147	.100	18000
	severe	.153	.096	18000
	average	.148	.092	54000
7	none	.171	.090	18000
	moderate	.160	.087	18000
	severe	.152	.086	18000
	average	.161	.088	54000

[a]Average across all levels of skew.

Table 6

Eta-squared Values for the Effects of Number of Items with Correlated Errors and Magnitude of Correlated Errors on the Difference between Alpha and True Reliability Estimates

		Sum of Squares	η^2
Repeated-measures effects			
	Reliability[a]	6574.933	0.403
	Reliability*Number of items with correlated errors	79.087	0.005
	Reliability*Magnitude of correlated errors	73.536	0.005
	Reliability*Magnitude*Number of items	17.806	0.001
	Error	4060.671	
	Total	10806.033	
Between-measures effects			
	Number of items with correlated errors	3.830	0.000
	Magnitude of correlated errors	3.703	0.000
	Magnitude*Number of items	.860	0.000
	Error	5484.313	
	Total	5492.706	
Overall	Total	16298.74	

[a]Repeated-measures effect of the difference between alpha and true reliability estimates 1 and 2.

Table 7

Alpha and True Reliabilities by Item Correlations and Number of Correlated Items for Entirely Categorical Data Sets

Correlation	Number of items	*n*	Alpha		True reliability 1[a]		True Reliability 2[b]	
			M	*SD*	*M*	*SD*	*M*	*SD*
.1	3	80173	.758	.161	0.484	0.232	0.610	0.164
	5	80172	.767	.152	0.477	0.230	0.597	0.164
.3	3	80167	.767	.153	0.478	0.230	0.598	0.164
	5	80176	.793	.130	0.460	0.225	0.564	0.167

[a]The squared correlation between true continuous scores and final observed scores.
[b]The squared correlation between categorized true scores and final observed scores.

Table 8

Alpha and True Reliabilities by Factor Loadings for Entirely Categorical Data Sets

Factor loading	Alpha		True reliability 1[a]		True Reliability 2[b]	
	M	*SD*	*M*	*SD*	*M*	*SD*
.4	.603	.130	.273	.177	.448	.105
.6	.789	.072	.463	.174	.628	.130
.8	.910	.031	.681	.116	.692	.156

Note: The third true reliability estimate equals .656, .849, and .947 for factor loadings of .4, .6, and .8, respectively.
[a]The squared correlation between true continuous scores and final observed scores.
[b]The squared correlation between categorized true scores and final observed scores.

Table 9

Estimated Error Correlations by Number of Categories, Skew, and Factor Loadings for Partially Categorized Data Sets

		No skew		Moderate Skew		Severe Skew	
Number of categories	Factor loadings	*M*	*SD*	*M*	*SD*	*M*	*SD*
2	0.4	.000	.018	.006	.019	.008	.020
	0.6	.007	.019	.040	.021	.048	.025
	0.8	.059	.020	.158	.026	.186	.034
	Average[a]	.022	.032	.068	.069	.081	.081
3	0.4	.000	.020	.007	.020	.009	.021
	0.6	.002	.021	.047	.022	.058	.026
	0.8	.022	.020	.192	.027	.230	.036
	Average	.008	.022	.082	.083	.099	.099
4	0.4	.000	.021	.003	.020	.008	.020
	0.6	.003	.021	.028	.021	.050	.022
	0.8	.030	.020	.141	.024	.206	.027
	Average	.011	.025	.058	.064	.088	.088
5	0.4	.000	.021	.006	.021	.011	.021
	0.6	-.001	.020	.045	.023	.065	.027
	0.8	.008	.021	.190	.027	.254	.036
	Average	.002	.021	.080	.083	.110	.108
6	0.4	.000	.021	.008	.020	.008	.021
	0.6	.004	.021	.052	.023	.049	.023
	0.8	.042	.022	.216	.028	.209	.028
	Average	.015	.029	.092	.092	.089	.090
7	0.4	.000	.020	.005	.021	.009	.020
	0.6	.006	.021	.038	.021	.052	.024
	0.8	.051	.022	.171	.024	.216	.028
	Average	.019	.031	.072	.075	.092	.093

Note. n=16200, 3000 per condition.

[a]Average across factor loadings.

Table 10

Eta-squared Values for the Effects of Number of Categories, Skewness, and Factor Loadings on the Estimate of Correlated Errors

	Sum of Squares	η^2
Number of categories	3.374	.002
Skewness	191.403	.117
Factor loading	579.388	.354
Categories*Skew	10.247	.006
Categories*Loading	2.780	.002
Skew*Loadings	163.864	.100
Categories*Skew*Loadings	9.505	.006
Error	86.421	
Total	1637.604	

Table 11

Alpha and True Reliability by Skew and Factor Loadings for Partially Categorized Data Sets

		Alpha		True Reliability 1[a]	
Skew	Loading	*M*	*SD*	*M*	*SD*
None	0.4	.616	.028	.616	.025
	0.6	.820	.019	.822	.016
	0.8	.928	.012	.930	.010
	Average[a]	.788	.131	.790	.131
Moderate	0.4	.590	.032	.589	.027
	0.6	.801	.023	.796	.021
	0.8	.916	.016	.901	.024
	average	.769	.137	.762	.132
Severe	0.4	.573	.040	.576	.033
	0.6	.787	.031	.787	.024
	0.8	.906	.023	.898	.025
	average	.755	.141	.754	.136
Average[b]	0.4	.593	.038	.594	.033
	0.6	.803	.028	.802	.025
	0.8	.916	.020	.910	.025
	average	.771	.137	.768	.134

Note: n = 54000, 18000 for each condition.

Note: The third true reliability estimate equals .656, .849, and .947 for factor loadings of .4, .6, and .8, respectively.

[a]The squared correlation between categorized true scores and final observed scores.

Table 12

Eta-squared Values for the Effects of Skew, Factor Loadings, and Number of Categories on the Difference between Alpha and True Reliability Estimate 1

		Sum of Squares	η^2
Repeated-measures effects			
	Reliability[a]	765.987	0.033
	Reliability*Skew	85.600	0.004
	Reliability*Loading	206.595	0.009
	Reliability*Categories	75.870	0.003
	Reliability*Skew*Loading	56.708	0.002
	Reliability*Skew*Categories	109.062	0.005
	Reliability*Loading*Categories	74.658	0.003
	Reliability*Skew*Loading*Categories	57.828	0.002
	Error	2027.303	
	Total	3459.611	
Between-measures effects			
	Skew	746.210	0.032
	Loading	13477.125	0.577
	Skew*Loading	520.684	0.022
	Skew*Categories	14.510	0.001
	Loading*Categories	129.770	0.006
	Skew*Loading*Categories	52.964	0.002
	Error	25.846	
	Total	4932.120	
Overall	Total	23358.842	

[a]Repeated-measures effect of the difference between alpha and both true reliability estimates.

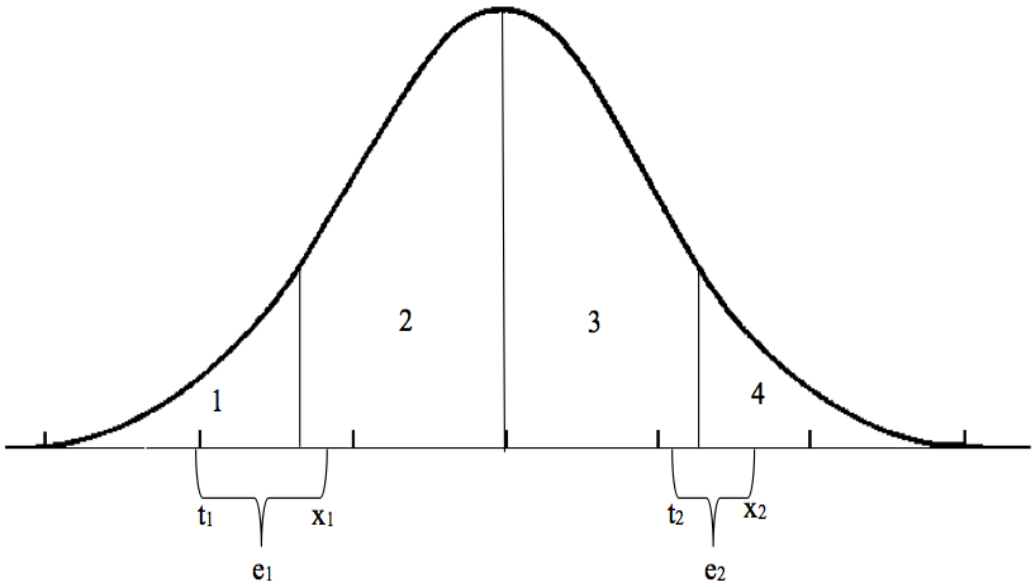

Figure 1. Illustration of misclassification error. True scores (t_x) should be categorized into a specific category; however the error component (e_x) causes the score to be classified differently.

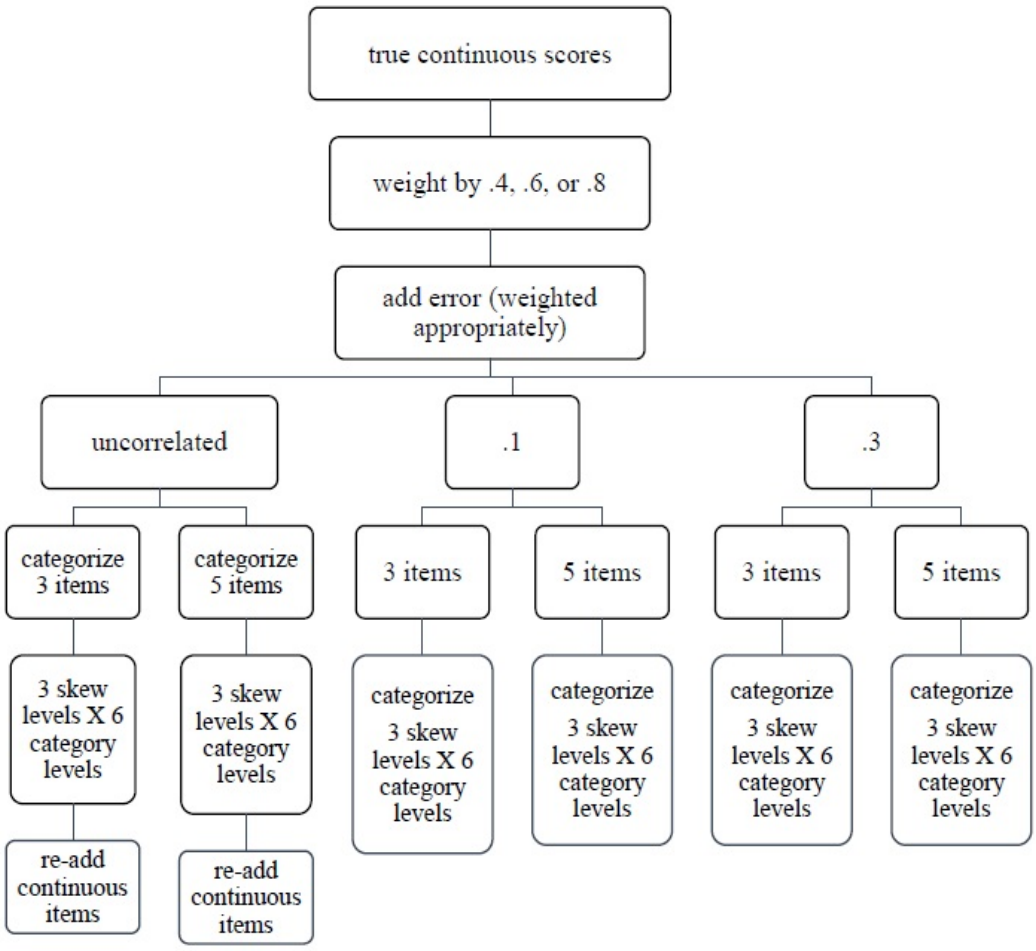

Figure 2. Data generation design.

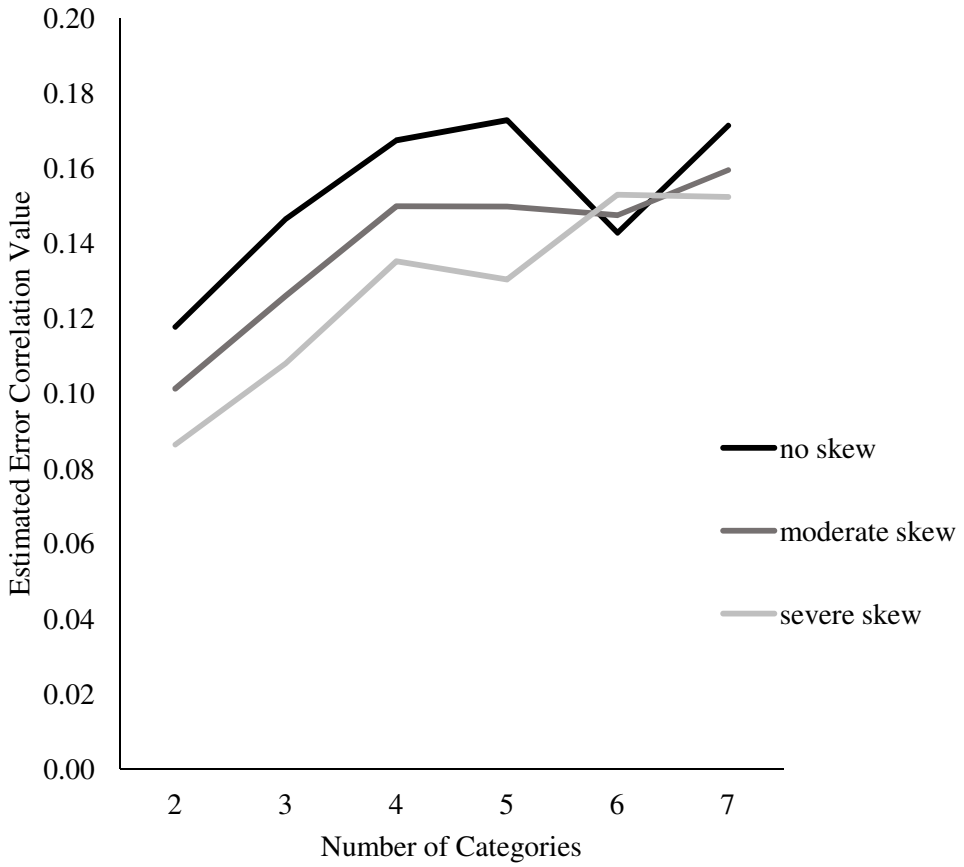

Figure 3. Average estimated error correlations as a function of skew and number of categories.

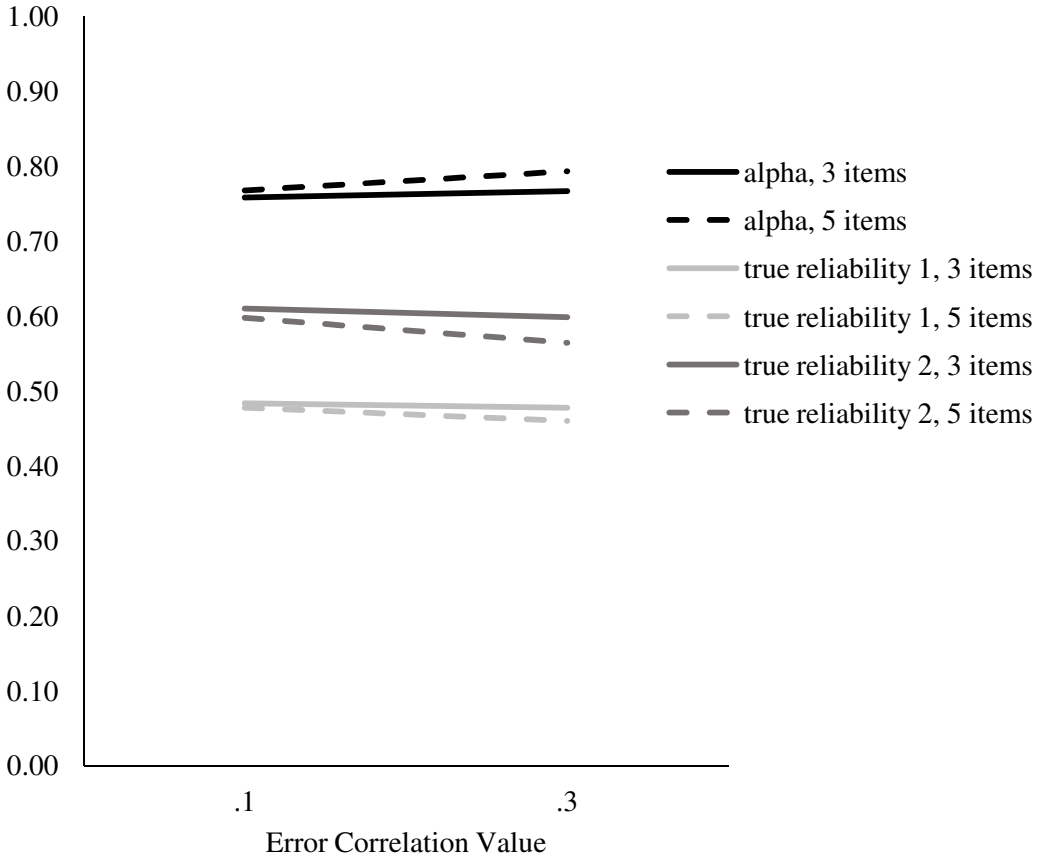

Figure 4. Alpha and true reliability as a function of the number of items with correlated errors and the magnitude of the correlated errors.

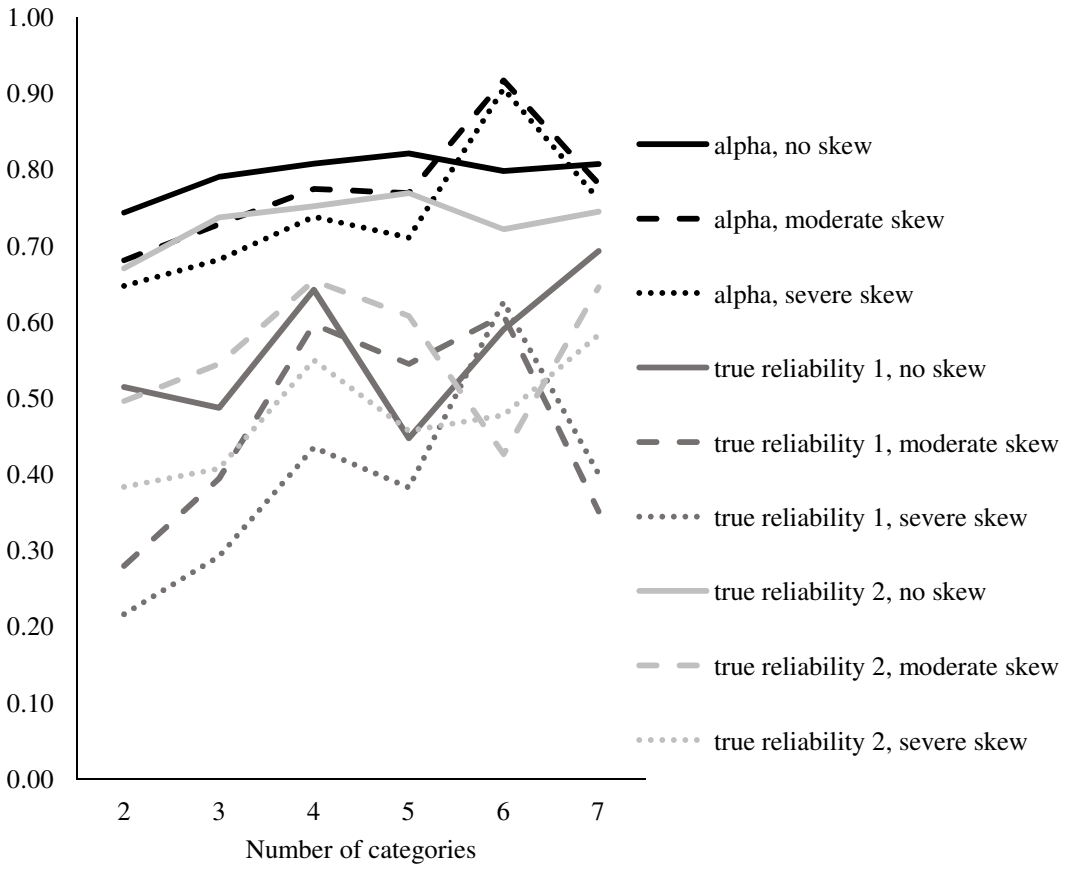

Figure 5. Alpha and true reliability as a function of skew and the number of categories.

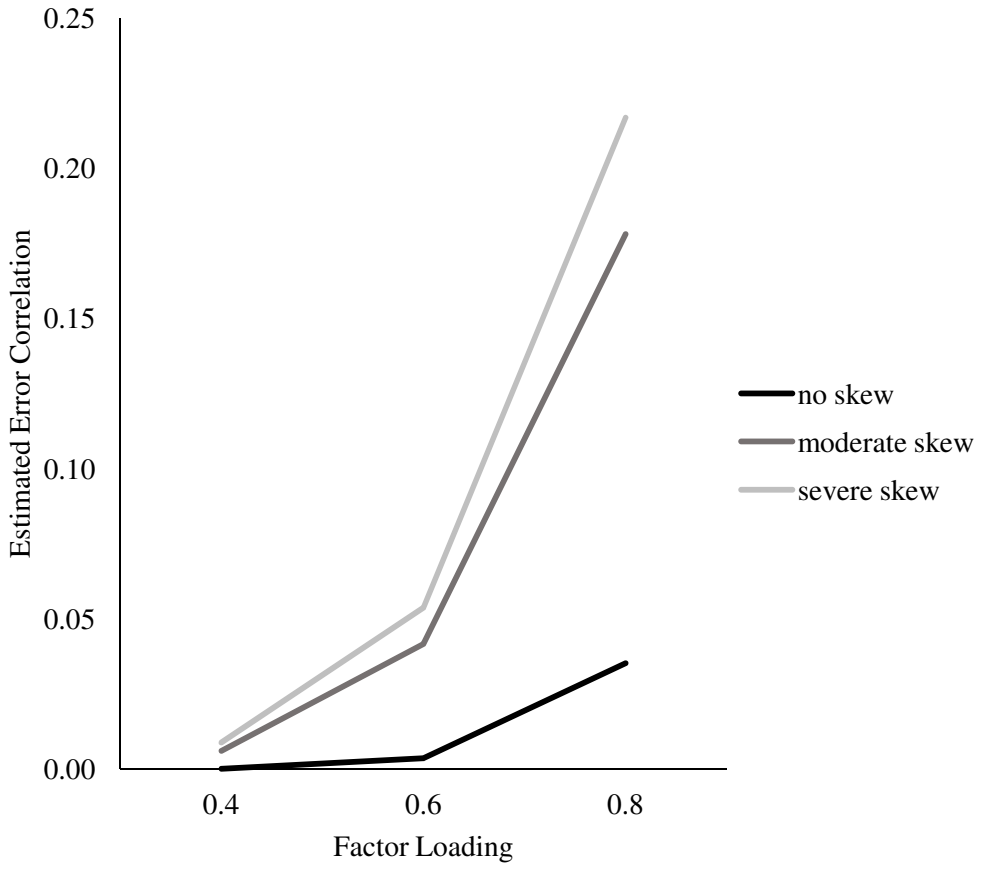

Figure 6. Average estimated error correlations by skew and factor loadings.

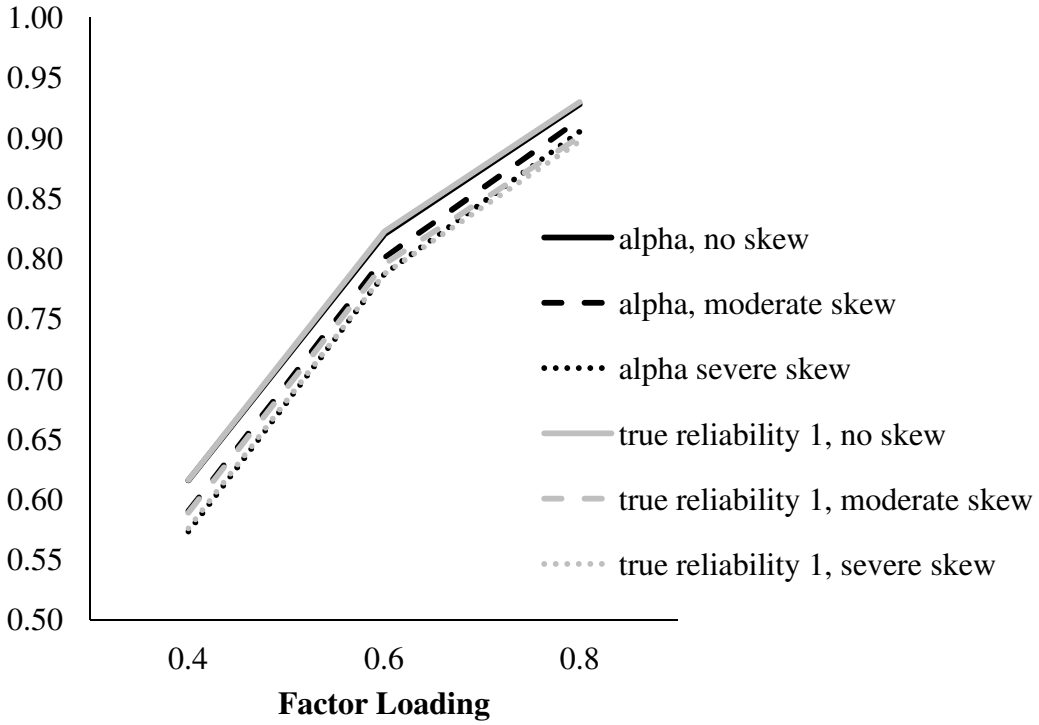

Figure 7. Alpha and true reliabilities as a function of loadings and skew. Note the change in y-axis range.