Portland State University

PDXScholar

Chemistry Faculty Publications and Presentations

Chemistry

7-6-2018

Moving Beyond Alpha: A Primer on Alternative Sources of Single-Administration Reliability Evidence for Quantitative Chemistry Education Research

Regis Komperda
Portland State University, regis.komperda@pdx.edu

Thomas C. Pentecost
Grand Valley State University

Jack Barbera
Portland State University, jbarbera@pdx.edu

Follow this and additional works at: https://pdxscholar.library.pdx.edu/chem_fac

 Part of the Chemistry Commons, and the Science and Mathematics Education Commons

Let us know how access to this document benefits you.

Citation Details

Published as Regis Komperda, Thomas Pentecost, and Jack Barbera, "Moving Beyond Alpha: A Primer on Alternative Sources of Single-Administration Reliability Evidence for Quantitative Chemistry Education Research," Journal of Chemical Education, 2018, 95, 9, 1477-1491, 35 DOI:10.1021/acs.jchemed.8b00220.

This Post-Print is brought to you for free and open access. It has been accepted for inclusion in Chemistry Faculty Publications and Presentations by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.

Moving beyond alpha: A primer on alternative sources of single-administration reliability evidence for quantitative chemistry education research

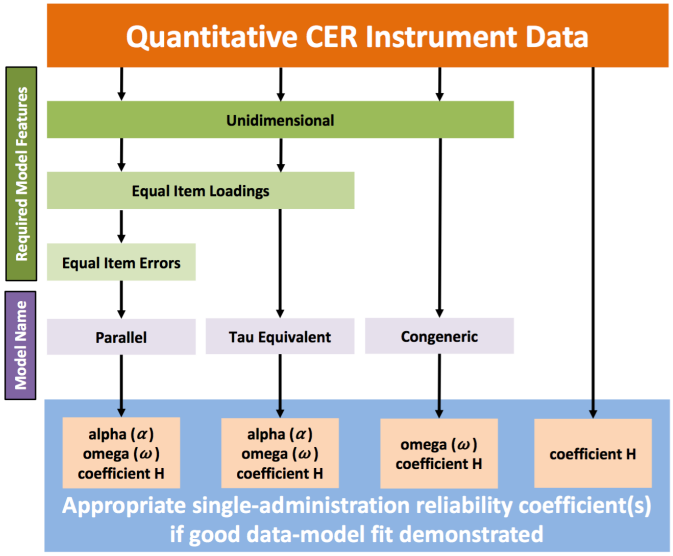
Regis Komperda[1], Thomas C. Pentecost[2], Jack Barbera[1]*

[1]Chemistry Department, Portland State University, Portland, OR.

[2]Chemistry Department, Grand Valley State University, Allendale, MI.

ABSTRACT

This methodological paper examines current conceptions of reliability in chemistry education research (CER) and provides recommendations for moving beyond the current reliance on reporting coefficient alpha (α) as reliability evidence without regard to its appropriateness for the research context. To help foster a better understanding of reliability and the assumptions that underlie reliability coefficients, reliability is first described from a conceptual framework, drawing on examples from measurement in the physical sciences; then classical test theory is used to frame a discussion of how reliability evidence for psychometric measurements is commonly examined in CER, primarily in the form of single-administration reliability coefficients. Following this more conceptual introduction to reliability, the paper transitions to a more mathematical treatment of reliability using a factor analysis framework with emphasis on the assumptions underlying coefficient alpha and other single-administration reliability coefficients, such as omega (ω) and coefficient H, which are recommended as successors to alpha in CER due to their more broad applicability to a variety of factor models. The factor analysis-based reliability discussion is accompanied by R code that demonstrates the mathematical relations underlying single-administration reliability coefficients and provides interested readers the opportunity to compute coefficients beyond alpha for their own data.



KEYWORDS
General Public; Chemical Education Research; Testing/Assessment; Chemometrics

Reference Information:
Regis Komperda, Thomas Pentecost, and Jack Barbera, "Moving Beyond Alpha: A Primer on Alternative Sources of Single-Administration Reliability Evidence for Quantitative Chemistry Education Research," *Journal of Chemical Education*, 2018, 95, 9, 1477-1491, DOI:10.1021/acs.jchemed.8b00220.

INTRODUCTION

Conducting any type of research relies on having high quality instrumentation available to measure the system under investigation. In quantitative chemistry education research (CER), the instrument is often something completed by research subjects to provide information about individuals or groups on one or more variables of interest, e.g. motivation, attitude, content knowledge, or misconceptions.[1,2] These variables are often used in CER to gauge the impact of a pedagogical reform or innovation. Development and evaluation of these instruments, also called tests, assessments, scales, surveys, inventories, or questionnaires, is the primary focus of the field of psychometrics. Determining whether or not an instrument provides high quality measurements relies on the collection and interpretation of psychometric evidence.

When describing the type of psychometric evidence reported in the *Journal of Chemical Education*, Arjoon, Xu and Lewis[1] used the *Standards for Educational and Psychological Testing*[3] as a framework to explore evidence reported for two key aspects of measurement quality, validity and reliability. Validity and reliability are interrelated aspects of psychometric measurement quality that have analogs in physical measurements: reliability describes the precision of a measurement, validity describes its accuracy. Arjoon et al. found that validity evidence was more widely reported within CER than reliability evidence. Despite the variety of ways to assess reliability given in the *Standards*, Arjoon et al. found only reliability coefficients were reported in the reviewed CER literature, specifically test-retest and Cronbach's alpha (α). For the 20 instruments reviewed by Arjoon et al., alpha was reported for 13 instruments while test-retest was reported for seven instruments and the intra-class correlation coefficient was reported for two instruments.

The prevalence of reporting alpha for instruments published in the *Journal of Chemical Education* from 2002–2011, as documented by Arjoon, Xu and Lewis,[1] persists in a current review conducted of 59 research articles in the *Journal* from 2012–2017 that used either a cognitive or affective instrument, shown in Table 1. The values in Table 1 sum to more than 59 because some articles reported multiple reliability measures. Of the 59 articles reviewed in which a psychometric instrument was used, roughly half (31) reported a value of alpha calculated from their own data and an additional three articles provided literature alpha values from previous uses of the instrument. After alpha, the

next most commonly reported reliability measure was test-retest, though this was only used in five studies. A separate study demonstrated an alternative to traditional test-retest approaches, the zeta-range estimator.[4] And a single study used the Kuder-Richardson KR 21 formula for dichotomously scored items.

Other measures of reliability reported included interrater reliability (not listed in Table 1) and values obtained from applying the Rasch measurement model. Reliability values from the Rasch model, a form of item response theory, are related to measures of reliability from classical test theory, including coefficient alpha, although the relation is complex.[5,6] Because interrater forms of reliability are derived from comparing coding done by raters,[7,8] not an individuals' instrument scores, they will not be further addressed in this discussion. It is concerning to note that in this review, over a quarter of the examined research articles (18) did not report any measure of reliability for their chosen instrument. It is unclear if this is a result of not conducting an examination of reliability or choosing not to report reliability information.

Table 1. Reliability Measures of Psychometric Instruments in *JCE* Research Articles 2012-2017

Reliability measure	Number of articles reporting
Coefficient alpha: Calculated	31
None	18
Test-retest	5
Rasch-derived	5
Coefficient alpha: From literature	3
Zeta-range estimator	1
Kuder–Richardson: KR 21	1

Digging more specifically into how alpha is described within CER literature, an examination of the previously described 31 studies in the *Journal* reporting alpha revealed that half of the studies described alpha as providing a measure of internal consistency, a finding echoed in the broader science education literature[9] and consistent with the way that alpha is described in the *Standards*[3] and much of the psychometric literature.[10–12] This leads to the question of what exactly is meant by internal consistency and why it is a desirable property. One interpretation seen in both the *Journal* articles and other education and psychology literature is that internal consistency indicates that all of the items are measuring the same underlying variable, sometimes referred to as unidimensionality or

homogeneity, although the equivalence of these terms is contentious.[13–17] While it may be beneficial to know an instrument is measuring a single variable of interest, that is, that the instrument is unidimensional, it has been demonstrated that alpha does not provide this information[13,17–19] and it is also not straightforward to see how internal consistency aligns with the idea of reliability as describing the precision of a measurement.

Confusion over the information provided by alpha, and an even more basic lack of familiarity with the term itself, were also apparent in a national survey of 1,436 chemistry faculty conducted in 2009–2010.[20,21] The study found that faculty felt fairly familiar with the term "Assessment Reliability," providing a median rating of 4 on a 5-point scale, corresponding to "I have heard this term before and have a sense of what it means." Yet, when assessing their familiarity with "Cronbach Alpha," the only term specific to reliability that was listed, the overall median rating dropped to 1 indicating "I have never heard this term before." The chemistry education subgroup of responses, representing 10% of the faculty, gave a median rating of 2 corresponding to "I have heard this term before but do not know what it means." This low level of understanding is puzzling given the high prevalence of alpha in CER literature.

What the study of faculty familiarity with assessment terminology and recent *Journal* literature review may indicate is that researchers and reviewers are aware that reliability evidence should be reported when collecting data from an instrument and also recognize that alpha is the most commonly reported type of reliability evidence.[9] Yet, there may be limited understanding of what the reported value of alpha is actually saying about the quality of the data obtained[9] or limited awareness of other alternatives to reporting alpha.[11,14,22] The confusion about what alpha represents is not limited to CER,[22,23] and the continued use of alpha is the subject of vigorous debate amongst psychometricians.[16,24,25] Even Cronbach had misgivings about the ubiquitous use of alpha saying, "I doubt whether coefficient alpha is the best way of judging the reliability of the instrument to which it is applied" (p. 393) and was embarrassed by alpha's association with his name given the formula's previous establishment in the psychometric literature.[26] Though Cronbach disliked both the label alpha[26] and its association with his own name, that is how the formula is commonly known

throughout the literature and therefore for the remainder of this paper it will be referred to as coefficient alpha, or simply alpha.

GOALS OF THIS RELIABILITY PRIMER

Given the ongoing discussions of reliability within the psychometric community, it is not the intention of this paper to attempt to resolve any of the debates over what alpha represents or whether it should ever be used. Instead, the current level of interest in alpha will be used as a starting point from which to provide an accessible overview of reliability and the assumptions underlying various methods of computing reliability coefficients for an audience more familiar with measurements of physical systems than the mathematical underpinnings of classical test theory and factor analysis. The first section begins by framing reliability in the context of physical measurements in order to demonstrate where the analogy between measuring physical systems and psychological variables begins to break down. Presentation of specific reliability formulas are avoided in this section in favor of focusing on the conceptual meaning of reliability. Interested readers are encouraged to consult any of the excellent sources available for a more mathematical treatment of reliability.[10,18,27–30]

Capitalizing on CER's embrace of factor analysis methods for examining the internal structure of instruments as a form of validity evidence,[1] a factor analysis approach to reliability is presented in the second section, along with alternatives to alpha based in factor analysis.[22,27,31–34] The factor analysis section is more technical and, though an overview of terms and notation is provided, some readers may find it helpful to consult other sources for an introduction to this methodology.[35–37] The intention of these presentations is to highlight the different ways reliability can be addressed, both conceptually and mathematically, and the limitations associated with each approach so that researchers have the information necessary to make an informed decision on how best to report and describe reliability in the context of their own research.

RELIABILITY: TRANSLATING PHYSICAL SCIENCE MEASUREMENTS TO PSYCHOMETRICS

Sources of Measurement Error

A critical component of making quality measurements in both the physical sciences and psychometrics is to identify and minimize the amount of error present. Due to the presence of error,

an observed measurement value represents an obscured version of the true value. In psychometrics, this is often stated formally with the expression

$$\text{True value} = \text{Observed value} - \text{Error} \quad (1)$$

All measurements have error associated with them; in those instances where the amount of error can be quantified, the true value of the measurement can be determined.

A frequent analogy used to bridge physical science and psychometrics, as mentioned previously, is that the accuracy of a measurement describes its validity while the precision of a measurement describes its reliability.[3,38] In the context of measurement error, validity is related to the amount of systematic error present while reliability is related to the amount of random error. To illustrate the difference between systematic and random error, consider the process of calibrating a thermometer using a known standard. In Figure 1, the known standard could represent the normal boiling point of water (99.974 °C).[39]

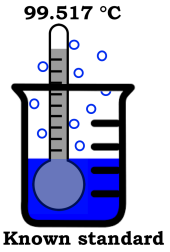


Figure 1. Calibration of thermometer with known standard to quantify systematic error

The difference between the true value of the known standard (99.974 °C) and the observed value (99.517 °C) can be entered into Equation 1, thereby giving the amount of error (0.457 °C), which can be used calibrate the thermometer. In this example, the error term represents systematic error.

While calibration of an instrument is often possible when making measurements of physical samples, there are more limitations when making measurements of psychological variables such as interest, motivation, or understanding. The primary issue is that it is never possible to measure a psychological variable directly and as a result, a true value can never be known and a known standard can never be established. Instead, response to a stimulus, such as an item on a test or survey, is measured and that measurement is used to draw conclusions about the underlying psychological

construct while the true amount of interest, motivation, or understanding is never knowable. In the context of Equation 1, this means that most psychological measurements provide only the observed value but no true value and therefore no way to identify the exact amount of systematic error present. However, even without calibration to a true value there are other types of evidence that can be obtained to help assess the validity of a psychological measurement.[1–3]

Though calibration addresses the amount of systematic measurement error present, it does not address random error. Consider using the calibrated thermometer from Figure 1 to measure the known standard multiple times under the same conditions; it is unlikely that the exact same values would be obtained each time (Figure 2). Instead, the spread in the resulting values provides information about the reliability of the thermometer data. A smaller spread in values indicates less random error and therefore greater reliability.

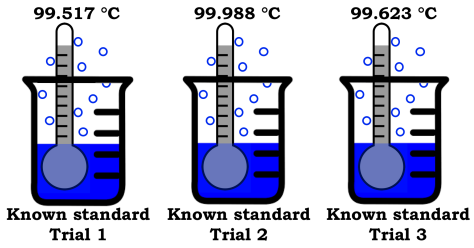


Figure 2. Multiple measurements with single thermometer to quantify random error

With multiple measurements of the same sample, shown in Figure 2, it becomes possible to calculate the standard deviation of the measurements (0.247) or the standard error of the mean (0.143), calculated as the standard deviation divided by the square root of the number of measurements, which can also be used to determine a 95% confidence interval for the mean [99.096; 100.323]. These three computations – standard deviation, standard error, and confidence interval – describe the consistency of the measurement, and are the typical types of precision used in physical measurements. This idea of consistency is echoed in psychometric descriptions of reliability. The *Standards* describe reliability as “the consistency of such measurements when the testing procedure is repeated on a population of individuals or groups” (p. 25).[3] This is analogous to the repeated measures depicted in Figure 2.

In comparison to physical samples, like those shown in Figure 2, the difficulty in measuring the same sample multiple times for psychometric research arises from limitations associated with having access to the same people on multiple occasions as well as the difficulty of identifying an appropriate interval for repeated measurements that does not result in test fatigue, recall of prior responses, or a change in the underlying variable being measured. In spite of these difficulties, a repeated measures design with two time points is sometimes undertaken and the resulting correlation between two sets of measurements is known as test-retest reliability. This process is illustrated in Figure 3 where the same thermometer is used to measure multiple samples at two time points. The samples measured in each trial in Figure 3 are no longer assumed to be at the same temperature, since there is no expectation that all samples within a population would have the same measured value. However, the measurement of each sample is assumed to be stable over time. If the measurements at both time points are very similar, the value of the correlation will be larger (closer to 1) and therefore test-retest reliability is said to be high. The correlation for the two sets of samples shown in Figure 3 is very high, 0.987, and this consistency can be seen in the plot of temperatures for each trial where all points are clustered near the line defining identical temperatures for each trial. Due to the temporal nature of these data, test-retest reliability has also been called the coefficient of stability.[7,40,41]

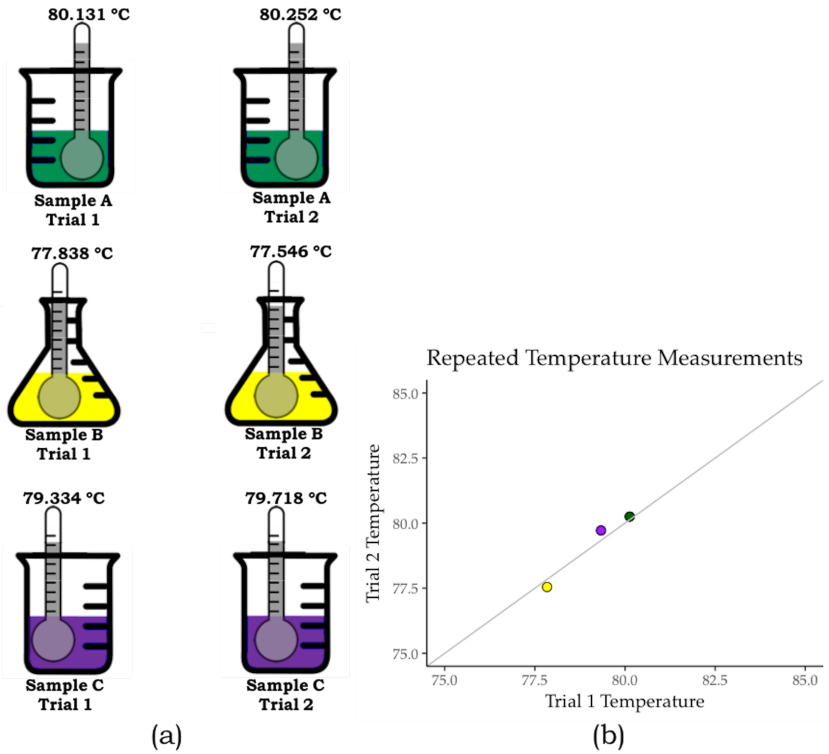


Figure 3. Measurement of multiple samples with differing temperatures at multiple time points with the same thermometer (a) plotted to show the similarity between sets of measurements taken at each time point relative to the line $y = x$ as an analogy for test-retest reliability (b)

The main assumption underlying test-retest reliability is that a high correlation between the two sets of measurements is a result of consistency in the underlying true values. It is therefore important to consider if the time interval between measurements supports this assumption. Considering the thermometer example, the time interval should not be so long that the set of samples has equilibrated to the ambient laboratory temperature thereby changing the true temperature value. In the same way for psychological measurements, the time interval should not be so long that learning or change in attitude has occurred. Additionally, while a shortened time interval is not typically a concern in a laboratory environment, for psychological measurements using too short a time interval may cause the person to remember his or her previous response, which could result in a strong relation between the measurements at the different time points that is not entirely due to the precision of the measurement. Details about factors to consider when using test-retest and stability coefficients to provide reliability evidence for psychometric measurements may be found elsewhwere.[42]

Reliability in Single-Administration Contexts

Given the difficulty associated with finding an appropriate interval for test-retest reliability, it is often easier to make measurements with multiple instruments simultaneously. In a laboratory context, this would be equivalent to measuring the same set of samples with multiple thermometers (Figure 4). Clearly, both instruments must be measuring the same variable so that comparison of the measurements is meaningful. In other words, it would not make sense to use a pH probe for a temperature measurement in the same way that it wouldn't make sense to include survey items about satisfaction with laboratory equipment when measuring test-taking anxiety.[9]

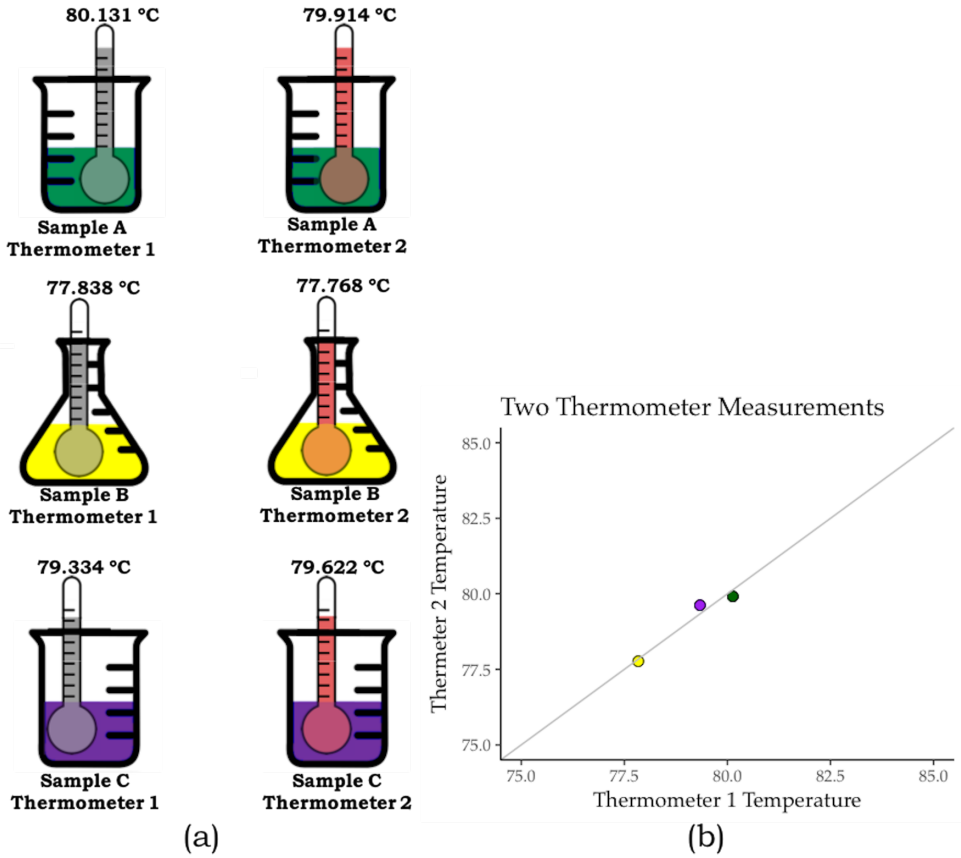


Figure 4. Measurement of multiple samples with differing temperatures using different thermometers simultaneously (a) plotted to show the similarity between sets of measurements taken with each thermometer relative to the line $y = x$ as an analogy for parallel-forms reliability (b)

As with test-retest reliability, a reliability value can be determined by finding the correlation between the data collected from the first thermometer and the data collected from the second thermometer (0.975). However, strength of this association no longer indicates consistency over time,

but rather consistency between the two thermometers. In psychometrics, this type of consistency is known as parallel- or alternate-forms reliability and is sometimes called the coefficient of equivalence.[7,40,41] If the two sets of measurements are taken on different days, this is known as the coefficient of stability and equivalence as it also incorporates aspects of test-retest reliability.[7,40,41] As with test-retest reliability there are some assumptions that must be met in order to conclude that a high correlation demonstrates high consistency of measurement. These assumptions primarily focus on ensuring the two instruments are making equivalent measurements, described as being parallel in a psychometric context. For the thermometers, each must use the same scale (e.g., Celsius, Fahrenheit, or Kelvin), they must produce the same average observed values with the same standard deviations,[7] and they must have the same amount of error.

To a chemist working in a laboratory setting, the situation described in Figure 4 is improbable in the context of making physical measurements as there are unlikely to be many situations where measurement of the same samples with two thermometers is preferable to repeated measurements by the same thermometer. However, for a chemistry education researcher in the context of conducting psychometric measurements, a single-administration approach to determining a reliability value is very often preferable to the logistics of setting up a test-retest condition. This preference for a single-administration approach to reliability in psychometric settings resulted in development of many well-known reliability coefficients, including coefficient alpha.

The reliability coefficients developed for single-administration contexts are generally known as internal consistency reliability, though concern exists over the usage and meaning of that label.[13,15,16,25] Using the analogy from Figure 4, this would involve using many thermometers (items) simultaneously on multiple samples (people). These single-administration reliability methods measure the strength of the relation between item responses and responses to the entire test or survey.[43]

One of the earliest single-administration reliability approaches, developed by Spearman and Brown,[44,45] was to take a test and divide it into two equivalent halves consisting of the same number of items. Then, reliability could be computed using formulas incorporating the correlations between each half of the test, referred to as split-halves reliability.[46,47] While the split-halves method alleviated concerns regarding the logistics of administering a test twice, there were other concerns with this

method. First, the parallel assumption, listed previously for the thermometers, must also hold for the two halves of the test. That is, the items must use the same measurement scale, they must result in the same average observed values and standard deviations,[7] and the amount of measurement error associated with each item must be the same and not related to the measurement error of the other items.[28] As with the thermometers, it can be difficult to identify sets of items that meet these strict assumptions. Second, it was unclear exactly how a test should be divided into halves. Different divisions of the test (e.g., even items vs. odd items, first half vs. second half) could result in different reliability values.[7]

To address these difficulties, additional single-administration reliability coefficients were developed that relaxed the parallel assumption and also allowed for computation methods beyond simply splitting the test in half. Guttman developed a series of six different single-administration reliability coefficients,[48] one of which (coefficient L_3) is mathematically equivalent to coefficient alpha.[27] Guttman first described these in 1945, which is the source of Cronbach's embarrassment that alpha came to be associated with him on the basis of his 1951 article.[26,49] Coefficient alpha is one such single-administration reliability estimate that relaxes the assumption of equal means and equal measurement errors for all items. However, the measurements for each item must be related to each other by an additive amount. In the thermometer context, this would be equivalent to using one Celsius thermometer and one Kelvin thermometer, but not a Fahrenheit thermometer because the size of each degree is different. Alpha also removes the debate over the possible ways to split a test and instead looks at the correlation between individual items and the overall test score. In this way, it represents the combination of all possible split-halves of the test.[7] Similarly, the Kuder Richardson (KR) 20 formula is mathematically equivalent to coefficient alpha in the case where items are dichotomously scored and KR 21 provides a simplification for dichotomously scored items of equal difficulties.[43,50]

Summary of Psychometric Reliability Coefficients

There are many types of reliability coefficients that can be calculated from data obtained using psychometric instruments. In light of the numerous options available, the selection of a reliability coefficient should be chosen to align with the goals of the research and the characteristics of the data

obtained from the instrument. For example, in situations where temporal stability is important, test-retest reliability may provide better information about reliability than a single-administration reliability estimate. In considering the characteristics of the data, alpha has some underlying mathematical assumptions about the relation between each item on the instrument (observed value) and the underlying variable being measured (true value). If these assumptions are not met, alpha provides a biased estimate of reliability.[24] Unfortunately, the most commonly used statistical software for computing alpha does not test to see if these assumptions have been met. On the other hand, adopting a factor analysis based approach to reliability will not automate the testing of these assumptions, but it does provide an opportunity to use the structure of the instrument to determine the appropriate approach for evaluating reliability from a single administration. A point to emphasize is that it is impossible to determine beforehand if a set of data will satisfy the assumptions necessary for the use of alpha. The following section describes how to test the assumptions in a factor analysis framework and depicts the relation between factor analysis and the previously described conceptions of reliability. A factor analysis approach to reliability is recommended both because CER is moving toward doing instrument development and testing in a factor analysis framework[1] and because factor analysis approaches take advantage of the more sophisticated computational methods available while relaxing some of the restrictive assumptions underlying coefficient alpha.

A FACTOR ANALYSIS APPROACH TO RELIABLITY

Visualizing Reliability as the Relation Between True and Observed Variance

Factor analysis methods such as exploratory factor analysis (EFA), confirmatory factor analysis (CFA), and structural equation modeling (SEM) are frequently used in CER to examine the internal structure of instruments as one method of providing validity evidence[1,51–54] and also provide a methodology for understanding relations among variables,[55–59] and examining group differences on variables of interest.[60,61] This section provides a brief overview of the notation conventions and conceptual underpinnings of factor analysis as they relate to reliability.[13,16,24,28,62] A more comprehensive introduction to factor analysis terminology and methodologies, including EFA, CFA, SEM, can be found in other sources.[35–37] In addition to the functionality of factor analysis as a tool for establishing validity,[1,3] it also provides a useful lens for understanding various aspects of single-

administration measures of reliability, including selection of which coefficient is most appropriate to report.[51,52] Though the transition to a factor analysis framework involves more complexity than the conceptual discussions of reliability used in the previous sections, factor analysis provides a concrete and tangible way to express and understand the underlying mathematical assumptions of alpha and other single-administration reliability coefficients. This section uses a theoretical and mathematical description of reliability framed in the context of single-factor models to illustrate how the different single-administration reliability coefficients are related, while each utilizing slightly different assumptions about the underlying structure of the data.

Though many software packages are available to conduct factor analysis, the statistical software R[63] is a free alternative that can perform factor analysis and other common instrument analyses, including computation of reliability coefficients.[64] The scenarios discussed in this section highlight mathematical relations that are critical for understanding the information provided by reliability coefficients. R code is provided in the Supporting Information to accompany each scenario. By providing the code necessary for readers to test these scenarios with simulated data, it is hoped that reliability coefficients become more tangible rather than feeling like a result of algorithms acting inside a mysterious black box. In addition, the R code provided can be modified to explore the scenarios presented here with other datasets and also used to calculate reliability coefficients for different datasets, including the readers' own.

In a broad sense, one goal of factor analysis techniques is to use observed relations among measured variables (i.e., correlations and covariances) to identify and model relations among underlying unobserved variables. These unobserved variables of interest are commonly known as factors or latent variables and are generally aligned with specific constructs that are being measured such as attitudes or motivation. One critical component is that the models partition the common variance of the factor, representing the true value, from the unique error variance associated with the observed variables. In this way, an analogy can be made between factor analysis and the formal statement of measurement error described previously in Equation 1.

In visual notations of factor analysis, measured variables are conventionally represented by squares while unobserved variables are represented by circles or ovals. Considering the three variables

in Equation 1, only the observed value is measured directly while both the true value and the error are not directly observable in situations where calibration with a known standard is impossible, which includes most psychological measurements. Figure 5 shows the variables in Equation 1 in factor analysis notation.

Figure 5. True value equation variables in factor analysis notation

Another important aspect of factor analysis is that relations among variables are depicted using an arrow notation where a single-headed arrow represents a causal relation between two variables. In the case of the three variables represented in Figure 5, the causal relation between the true value, observed value, and error is more apparent after rearranging Equation 1 to solve for the observed value.

$$\text{Observed value} = \text{True value} + \text{Error} \tag{2}$$

Rearranged as Equation 2, it is clear that the observed value of a measurement is the aggregate of some amount of true value along with some amount of error. This mathematical relation can be depicted in a factor analysis model with the addition of two arrows pointing toward the observed value, one from the true value and one from the error. Figure 6 describes a situation where a single item is used as a measurement. The values of 1 over each arrow are implicit weights for the true value and error in Equation 2, meaning that the true value and error each contribute their full amount into the observed value and also that the units of the true value and error are the same as the observed value. Additionally, the lack of direct connection (i.e., an arrow) between the true value and error in Figure 6 indicates their independence from one another.

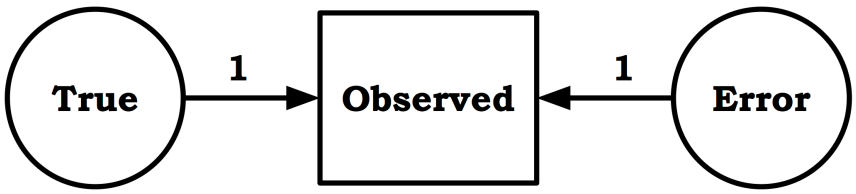


Figure 6. True value equation in factor analysis notation with causal arrows

When measurements of multiple subjects are made with that single item, there is now variance associated with the observed value. Factor analysis can be used to model how much of the observed variance is error variance and how much is variance of the true value. Using σ^2 to denote variance, Equation 2 can be rewritten using the variance of each term:

$$\sigma_O^2 = \sigma_T^2 + \sigma_E^2 \quad (3)$$

Considering the earlier description of reliability as the precision of a measurement, having more of the observed variance (σ_O^2) be due to true variance (σ_T^2) than error variance (σ_E^2) would indicate a more reliable measurement. This leads to one of the most common mathematical descriptions of the reliability coefficient, the ratio of true variance to observed variance, shown in Equation 4, which can be algebraically rearranged using the equality from Equation 3.[18,43] Note that Equation 4 uses the correlation symbol for reliability (ρ_r), in line with how reliability can be described as the correlation between sets of values, or more specifically, the squared correlation between observed and true values.[27,30] This reliability coefficient, ρ_r, represents a theoretical conception of reliability, not any specific type of reliability coefficient. Its relation to coefficient alpha, also referred to simply as alpha, will be described shortly.

$$\rho_r = \frac{\sigma_T^2}{\sigma_O^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} = \frac{\sigma_O^2 - \sigma_E^2}{\sigma_O^2} = \rho_{OT}^2 \quad (4)$$

Defined in this way, as the amount of true variance (σ_T^2) increases, and therefore the amount of error variance (σ_E^2) decreases, the reliability value increases. The reliability value ranges from 0 to 1, where a value of 1 indicates all of the observed score variance is true score variance.

Though this mathematical definition of reliability may be conceptually helpful, the practical difficulty is that in the context of psychological measurements, the true value, and therefore true variance, is never knowable. Alpha and other reliability coefficients offer some methods to circumvent this problem, but they come at the cost of requiring strong assumptions. In a factor analysis driven approach, these assumptions can be tested, and in some cases, avoided.

Visualizing Assumptions Underlying Alpha

In situations where more than three items are used to measure a single underlying latent variable, factor analysis provides an estimation of how much of the observed variance is due to the variance of a common construct of interest that influences true values and how much is due to the error variance of the items. Figure 7a shows a simplified factor model for a four-item instrument where the true values for all four items (A-D) are influenced by the common construct the items are intended to measure,[32] such as self-efficacy. The true values also have some amount of residual variance unexplained by the common construct, omitted from the model for simplicity. The model in Figure 7a can be simplified further to the model in Figure 7b where the true values are no longer explicitly shown and the error terms now represent a combination of both individual item error as well as the residual unexplained variance of the true value. More detailed explanation of these simplified models can be found in the interactive online tutorial module[65] and accompanying article[32] by Hancock and An. These models can be used to consider a wide variety of possible relations between the common construct and the observed responses to the items, assuming the common construct has been standardized by setting its variance to one.

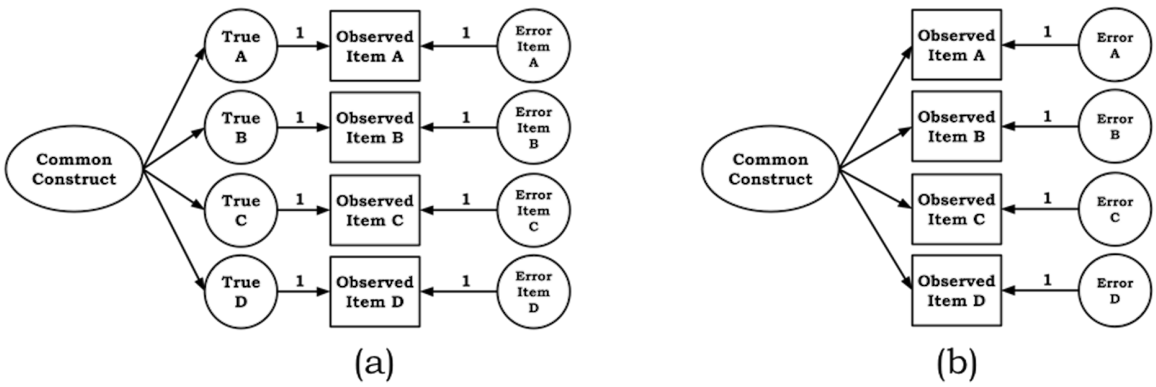


Figure 7. Factor model for four-item instrument showing true values for each item, omitting residual unexplained true value variance terms (a), and simplified to show only the common construct with standardized variance assumed (b)

In the following sections, some of these possible relations, and their factor models, will be used to illustrate methods for testing assumptions that underlie alpha. For each case, the appropriateness of coefficient alpha as a measure of reliability will be described. A factor model will also be used to illustrate alternatives to alpha for situations when the assumptions that underlie alpha are not met. R code in the Supporting Information is provided for readers who wish take a hands-on approach to exploring the relationship between different factor models and alpha using the provided simulated data or by importing their own data into R.

The Parallel Model: Identical Item Properties

Considering the four-item model in Figure 7, one possible relation between each observed value and the common construct is that the common construct is related to all of the items to the same degree. In the model, this would be equivalent to setting the value of all relations between the common construct and the observed items, known as loadings, equal.[13,28,32] In Figure 8, this equality is indicated by assigning the same value, denoted as λ, to each loading. This is the same restriction described previously in the discussion of parallel-forms reliability, now focused on parallel items. The same underlying assumptions apply here, such that the items must be measured on the same scale, the items must have same relation with the common construct, the error term of one item must not be related to the error term of any other item, and the items must have the same amount of error. This last assumption is described as having equal error variances in a factor analysis framework. In Figure

8 this assumption is indicated by assigning the same name to each error term, though this is a simplified depiction that is described more fully elsewhere.[32]

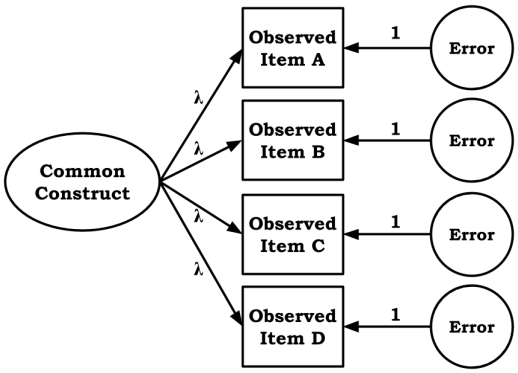


Figure 8. Parallel four-item model assuming a standardized common construct

When the conditions of the parallel model in Figure 8 are met, the value of alpha is equal to the squared correlation between the common construct and a composite score computed by summing responses to the four items.[22,27] According to this relation, larger alpha values indicate a stronger association between the common construct and the composite score, and therefore less error in the composite score. Though the value of the common construct is never known when working with real data, simulated data can be used to demonstrate this mathematical relation. The Supporting Information provides R code that can be used to generate simulated data following the model in Figure 8. These data can then be used to confirm the mathematical relation between alpha and the correlation between the common construct and composite score.

The Tau Equivalent and Essentially Tau Equivalent Models: Unequal Item Errors

The parallel model represents a highly restrictive set of conditions for the observed items unlikely to be met in most research settings. A less restrictive model, known as the tau equivalent model relaxes the restriction of having equal amounts of error associated with each item. This model is shown in Figure 9, where each item now has its own unique error term. As with the parallel model, simulated data can be used to show that under these conditions, alpha is still equal to the squared correlation between the common construct and the composite score.

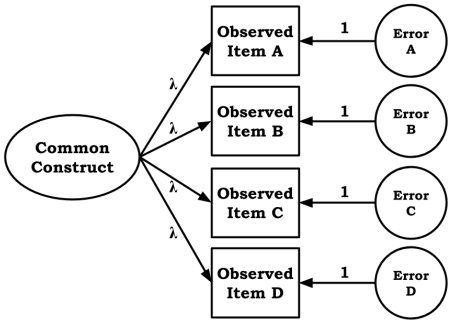


Figure 9. Tau equivalent four-item model assuming a standardized common construct

The restrictions of the tau equivalent model can be relaxed even further by allowing the relation between the common construct and the observed values to differ by an additive constant while maintaining the restriction of equal loadings. This model is known as the essentially tau equivalent model. In the thermometer analogy, this was described as using thermometers in both Celsius and Kelvin where the degree size is the same, but the scales differ by an additive constant. The equivalent degree size is what allows the relation between the observed and common construct to maintain the same linear relation describe in the factor analysis context by the loading values. Visually, the factor model for the essentially tau equivalent model is identical to the tau equivalent model shown in Figure 9. As with the parallel and tau equivalent models, simulated data can be used to show that under essentially tau equivalent conditions, alpha is still equal to the squared correlation between the common construct and the computed composite score.

The Congeneric Model: Unequal Item Errors and Unequal Relations with the Common Construct

Relaxing the restriction of equal loadings describes a congeneric model, represented in Figure 10 by assigning a unique value to each loading.

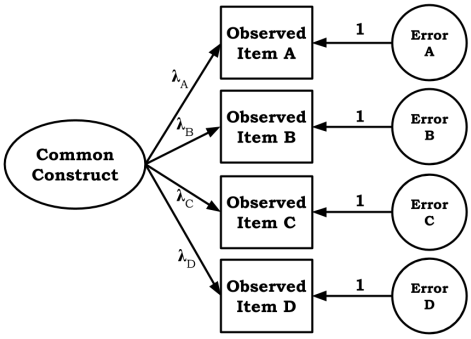


Figure 10. Congeneric four-item model assuming a standardized common construct

Under these conditions, each item is no longer restricted to have the same degree of association with the common construct. The congeneric model represents a typical situation encountered in psychometric research, including within CER as can be seen when researchers publish their loading values for factor models.[53–55,57,59–61] In the context of the thermometer example, this would allow thermometers of any scale to be used. However, under congeneric conditions, alpha is no longer equal to the squared correlation between the common construct and the computed composite score. When each item no longer has the same degree of association with the common construct, alpha is actually less than the squared correlation between the common construct and the computed composite score. This lowering of alpha relative to the squared correlation is where descriptions of alpha as the lower bound of reliability are derived from,[16] though if the item errors are not independent alpha can also overestimate reliability.[66] Again, the simulated data in the Supporting Information can be used to confirm this relation.

Relation Between Alpha and Factor Model Features

The different factor model conditions presented in Figures 8 through 10 demonstrate that alpha is only equal to the conceptual idea of reliability as a correlation between common construct and composite scores when the loadings are equal as in the parallel, tau equivalent, and essentially tau equivalent models. Additionally, all models have assumed no relation between the errors associated with each item. The presence of these types of correlated errors is known to bias alpha.[66,67] Therefore, before reporting a value of alpha, it would be appropriate to consider whether or not the data meet the assumptions of parallel, tau equivalent, or essentially tau equivalent models. If these assumptions are

not met, other alternatives for determining reliability are available and these alternatives will be discussed in the following sections.

Relation Between Unidimensionality and Reliability

The previously described relations between the common construct influencing the true values for each item and composite scores only hold for single-factor models like the ones shown in Figures 7 through 10. As numerous studies have demonstrated, alpha is not an index of unidimensionality;[13,18,19] unidimensionality is a requirement for the value of alpha to be meaningful.[9,22] If a multifactor model exists, it is more appropriate to report alpha for each individual subscale. In the words of Cronbach, "tests divisible into distinct subtests should be so divided before using the formula" (p. 297).[49] As will be briefly discussed later, other reliability coefficients exist that are more applicable to multifactor situations.[13]

Testing the fit between the data and the different single-factor model features to determine if the assumptions for reporting alpha have been met can be done through the use of confirmatory factor analysis (CFA). Numerous resources exist to explain how to conduct and evaluate the results of CFA,[35,37,68] and some sample R code for conducting CFA is provided in the Supporting Information. While performing CFA, it is necessary to be aware of important considerations such as sample size requirements,[69] and data characteristics, including missing data and whether data are continuous and normally distributed,[68] though methods exist for dealing with these situations.[70,71]

Alternatives to Alpha: McDonald's Omega for Composite Scores and Coefficient H for Weighted Composite Scores

If CFA indicates good data-model fit to parallel, tau equivalent, or essentially tau equivalent single-factor models, alpha may make sense as a way to report reliability. Even beyond demonstration of meeting the assumptions underlying alpha, reporting CFA results is helpful to the broader CER community since the information provided by CFA also support aspects of validity.[1,3,9] If CFA demonstrates that the data do not fit the more restrictive parallel, tau equivalent, and essentially tau equivalent models with equal loadings but instead show good fit to a congeneric model, there are other reliability coefficients that can be used instead of alpha. Of these, McDonald's omega (ω) is the most conceptually similar to alpha.[22,27,31,32] As shown in Equation 5, McDonald's omega is analogous to Equation 4 in defining reliability as the ratio of common construct variance to total variance. In

Equation 5, assuming a standardized common construct, the squared sum of the item loadings $(\sum \lambda)^2$ describes the common construct variance as the amount of observed variance explained by the underlying common construct, and total variance is the combination of the squared sum of the item loadings and the sum of the error variances, represented as $\sum \theta$. Omega, like alpha, ranges from 0 to 1 where 1 indicates that all of observed variance is the common factor variance.

$$\omega = \frac{(\sum \lambda)^2}{(\sum \lambda)^2 + \sum \theta} \quad (5)$$

Under congeneric model conditions, omega will be greater than alpha. When the omega formula is applied to models meeting the parallel, tau equivalent, or essentially tau equivalent assumptions, omega is equivalent to alpha.[22,27] This property of omega can be confirmed using the R code provided in the Supporting Information to calculate omega for the previously described scenarios. The fact that omega and alpha are identical when the assumptions for alpha are met means that it may be prudent simply to report omega when instrument data show good fit to a single-factor model. However, some examples exist in the CER literature of providing alpha when the assumption of equal loadings is met, but otherwise reporting omega for single-factor models showing good data-model fit.[51,52]

Alpha and omega are designed to address reliability of composite scale scores calculated as simple sums or averages of individual items, however in some situations it may make more sense to calculate an optimally weighted scale score. Weighted scale scores are particularly appropriate when items have a wide range of values for loadings on a single factor. When computing optimally weighted scale scores, items with stronger relations to the factor are weighted more heavily in the composite score than items with lower loadings.[22] In these situations, coefficient H[33,34] may provide a more meaningful indicator of reliability. Coefficient H also ranges from 0 to 1 and is computed using the values of the item loadings where a larger value of coefficient H indicates larger average loadings of the items. Because it provides an index of how strongly the measured items are related to the latent construct of interest, coefficient H has also been described as construct reliability,[34] as well as construct replicability and maximal reliability.[72] Like omega, coefficient H is equal to alpha when the assumptions of a parallel, tau equivalent, or essentially tau equivalent model are met.

Computing Alternatives to Alpha in R

Both omega and coefficient H can be calculated by hand after conducting CFA and obtaining values for loadings and error variances; methods also exist for using Mplus to compute omega,[32] and functions exist within R to automatically calculate omega and coefficient H from raw data. Within R, multiple packages provide functions for computing reliability coefficients, including userfriendlyscience,[73] MBESS,[74] semTools,[75] and coefficientalpha.[76] The function scaleStructure()[23,77] found in the package userfriendlyscience is recommended both for ease of use and for the variety of single-administration reliability coefficients it provides. The R code in the Supporting Information describes how to download and install this package and use the function scaleStructure() to compute single-administration reliability coefficients from raw data.

Figure 11 shows partial output from the scaleStructure() function which includes alpha, omega, coefficient H, and other reliability coefficients that are described briefly here and more fully by McNeish.[22] The scaleStructure() output also provides bootstrapped confidence intervals for alpha and omega[78] as well as estimates of ordinal alpha and ordinal omega when the function detects that the raw data are categorical.[79] Examination of the output of scaleStructure() shows three different reliability values named omega.

```
Information about this analysis:

               Dataframe: parallel[, 1:4]
                   Items: all
            Observations: 10000
  Positive correlations: 6 out of 6 (100%)

Estimates assuming interval level:

           Omega (total): 0.94
    Omega (hierarchical): 0.2
 Revelle's omega (total): 0.94
Greatest Lower Bound (GLB): 0.94
           Coefficient H: 0.94
        Cronbach's alpha: 0.94
Confidence intervals:
           Omega (total): [0.94, 0.94]
        Cronbach's alpha: [0.94, 0.94]
```

(a)

```
Information about this analysis:

               Dataframe: congeneric[, 1:4]
                   Items: all
            Observations: 10000
  Positive correlations: 6 out of 6 (100%)

Estimates assuming interval level:

           Omega (total): 0.84
    Omega (hierarchical): 0.78
 Revelle's omega (total): 0.8
Greatest Lower Bound (GLB): 0.85
           Coefficient H: 0.87
        Cronbach's alpha: 0.77
Confidence intervals:
           Omega (total): [0.83, 0.84]
        Cronbach's alpha: [0.76, 0.78]
```

(b)

Figure 11. Partial output from function scaleStructure() for data fitting a parallel (a) and congeneric (b) model

McDonald's omega is the first line of the reliability coefficient output, called Omega (total). The next value, Omega (hierarchical)[31] is more applicable to models with general factors as well as specific factors within a multidimensional factor model. Since this situation describes a model not commonly used for instruments in CER, it will not be discussed further. More similar to McDonald's omega total is Revelle's omega (total)[18] which uses a more complex mathematical process for generating the factor solution, but also considers more complicated factor structures than unidimensional models. Again, this complexity is likely unnecessary for most typical CER. Finally, the Greatest Lower Bound (GLB)[19,80] refers to a suite of methods examining the covariance matrix for the items, but it tends to be biased for smaller samples sizes and currently cannot be computed for ordinal data.[22] For this reason, GLB is unlikely to be a useful option for most CER.

Though not all of the reliability coefficients provided by scaleStructure() are likely to be relevant for most CER, it is illustrative to see how different methods of defining and computing reliability can result in different calculated values depending on which factor model the data fit. A natural next question might be to wonder whether there is a specific numeric value that should be obtained in order to declare the data derived from an instrument or scale as reliable. Higher values of reliability are preferable; however the idea of a universally standard "acceptable"[9] reliability value, such as the commonly cited cutoff of 0.70, is a myth stemming from the incorrect interpretation of Nunnally.[10,81,82] Rather than worrying about meeting an arbitrary threshold for a reliability value, it is more important to consider which type of reliability value is most appropriate to report given the stakes of the assessment, the context of the measurements, the structure of the instrument, and characteristics of the data. Then, interpretation of the reliability value can be used to explain whether or not it is acceptable in that context for that intended use.[3] As summarized by Bandalos, "there is no substitute for thoroughly thinking through the context of testing and purposes to which the test will be put and for using these to guide decisions about values of reliability coefficients" (p. 184).[42]

Summary of Reliability in a Factor Analysis Framework

Though alpha, omega, and coefficient H can be calculated directly from raw data using R packages, it is strongly recommended that the assumptions for each reliability coefficient are checked with CFA before undertaking the reliability calculation. While internal factor structure can be examined with

exploratory factor analysis (EFA), EFA is not appropriate for analysis of an instrument that has a theoretical rationale for a specific factor structure.[83] Additionally, the validity evidence from EFA is weaker than CFA because EFA does not provide the data-model fit indices that allow for testing how well the data fit the proposed theoretical model, thereby limiting the ability to determine whether or not a single-factor model is appropriate for the data. The additional steps required to conduct CFA are worthwhile since the CFA results can be used to provide evidence for the validity of the measurement while simultaneously evaluating assumptions underlying the reliability evidence.

In some situations the calculated values of alpha, omega, and coefficient H may be very similar, but it would be incorrect to assume that this makes their use interchangeable.[84] Just as assumptions are checked and reported before using and reporting outcomes from other statistical methods such as ANOVA, similar rigor should be standard for reliability. Without first demonstrating that the data have good fit to a single-factor model, reporting of either alpha or omega is meaningless. The mathematical examples provided in the supporting R code and other sources[13,17−19] have demonstrated that alpha does not provide as good an estimate of reliability if the data do not fit a unidimensional model with uncorrelated errors and equal item loadings as in the parallel, tau equivalent, or essentially tau equivalent models. Similarly, it does not make sense to report a value of coefficient H for a latent variable if data do not show good fit to a factor model. In situations where an instrument is known to be composed of multiple scales where scores will be reported separately, each scale should be evaluated to determine if it fits a single-factor model, and a reliability value should be provided for each set of scale data.

Limitations of Reliability in a Factor Analysis Framework

While there are many benefits of using a factor analysis approach to psychometric data analysis, such as obtaining validity evidence in addition to reliability evidence, there are also difficulties associated with moving to this approach. First, it is likely that some of the popularity of alpha in CER has arisen due to the ease of obtaining it in software, like SSPS, frequently used for data analysis.[64,85] Even with the increasing user-friendliness of R and the availability of functions for computing additional reliability coefficients beyond alpha, moving away from alpha still represents an additional layer of difficulty for most researchers.

Paradoxically, in some ways reducing the activation energy required to compute additional reliability coefficients from raw data has the potential to create new issues with reporting of reliability coefficients, such as omega and coefficient H, in contexts where they are not applicable due to not meeting underlying statistical assumptions. For this reason, it is also important to recognize what information single-administration reliability coefficients do and do not provide. Omega and alpha only provide appropriate reliability measurements if the goal of reporting a reliability value is to say something about the relative proportion of common construct variance to total variance with data that fit a single-factor model. That is, when statistical assumptions are met alpha and omega can be used to estimate the amount of error present in the measurement, specifically when using an equally weighted composite score. The interpretation of coefficient H is different in that H can be interpreted as providing information about the quality of a construct, as defined by having stronger relations between the construct and its indicator variables. As coefficient H represents the reliability of a composite score obtained from an optimally weighted set of items it is also the maximum the reliability can be within a given sample. For all reliability coefficients discussed, there is no set target value at which reliability crosses a threshold to become acceptable. Instead, it is necessary to justify why values may be appropriate for a particular research context given properties of the instrument, the subjects, and the setting.

Another concern with the ease of computing single-administration reliability coefficients, including alpha, is that functions such as scaleStructure() do not evaluate the underlying assumptions of unidimensionality or show data-model fit as would be obtained from performing CFA. This means that acceptable values for reliability coefficients may be obtained even if data do not show good fit to a single-factor model. No single-administration reliability coefficient is a substitute for the data-model fit information provided by CFA; conversely good data-model fit is not a substitute for acceptable reliability values.[86] While there are many benefits to using a factor analysis framework to approach reliability, there are also limitations, primarily related to the larger sample size requirements and additional analysis steps required as well as a stronger emphasis on the theoretical framework underlying the instrument. However, the CER community has made great strides in embracing a more

rigorous approach to evaluating measurement quality,[1] and it is anticipated that these types of analyses will soon become standard procedure.

RECOMMENDATIONS FOR REPORTING RELIABLITY OF PSYCHOMETRIC MEASUREMENTS

Choosing the Appropriate Reliability Estimate

Reliability provides information about the precision associated with a measurement. In a psychometric context, this is often defined as the relative proportion of common construct value to the total observed value. Alpha is only one of many mathematical methods for computing this proportion in situations where the value of the common construct that influences true values for each item is unknown, as is the case with psychometric measurements. When choosing how to address the reliability of measurements made using psychometric instruments, it is important to remember that, similar to validity, there are a variety of types of reliability that can be reported. The type of reliability reported should be aligned with the research goals and type of data collected.

- Test-retest reliability provides information about variability over time
- Parallel- or alternate-forms reliability provides information about variability across items
- Single-administration reliability values, including alpha, McDonald's omega, and coefficient H, provide information about the relations between individual items and a composite score

Checking Assumptions and Prerequisites

Each type of reliability has its own benefits and limitations as well as underlying assumptions that should be met before the value is reported, just like any other statistical test. When the assumptions are not met, reliability values no longer provide accurate estimations of the amount of random measurement error present. Some of these assumptions, such as the mathematical relation between the observed value of each item and the value of the common construct of interest that underlies single-administration reliability coefficients, are best examined in a factor analysis framework. The single-administration reliability coefficients themselves do not provide information about unidimensionality or demonstrate that the items are measuring a single construct, only factor analysis can test for these characteristics of the data. The information provided by factor analysis is a necessary prerequisite for moving forward with reporting alpha, omega or coefficient H. The factors to

consider when deciding whether to report alpha, omega or coefficient H are summarized below and in Table 2.

- Alpha should only be reported for instruments or scales showing good data-model fit to a single-factor model where each item is associated with the common construct to the same degree (i.e., a parallel, tau equivalent, or essentially tau equivalent model)
- Omega relaxes the restrictions of alpha by allowing the items in the single-factor model to be associated with the common construct to different degrees (i.e., a congeneric model)
    - When the mathematical assumptions for alpha are met, alpha and omega are equivalent, making omega a more universally appropriate single-administration reliability value for most contexts
    - Both reliability coefficients are appropriate for situations where data will be used to calculate a simple composite score
- Coefficient H has no structural assumptions other than demonstrating that the instrument or scale has good data-model fit to any proposed model
    - When assumptions for the parallel, tau equivalent, or essentially tau equivalent model are met, coefficient H is equal to alpha and omega
    - Because coefficient H provides a summary of the strength of the relation between the items and the common construct, it is more appropriate for situations where data will be used to calculate a weighted composite score

Table 2. Data characteristics and appropriate reliability coefficients for each model

	Model Name		
Data Characteristics	Parallel	Tau Equivalent or Essentially Tau Equivalent	Congeneric
Unidimensional	Yes	Yes	Yes
Equal item loadings	Yes	Yes	No
Equal item error values	Yes	No	No
Reporting Reliability	Parallel	Tau Equivalent or Essentially Tau Equivalent	Congeneric
Appropriate single-administration reliability coefficient(s)	alpha (α) omega (ω) coefficient H	alpha (α) omega (ω) coefficient H	omega (ω) coefficient H

Though examining the internal structure of an instrument with factor analysis is a necessary step in the process of reporting a single-administration reliability coefficient, the internal structure of the instrument provides validity evidence, not reliability evidence. In situations where conducting factor analysis is not feasible due to small sample sizes, researchers are encouraged identify and use instruments that have strong evidence for single-administration reliability with populations similar to those in the small-sample research project. In those situations, it is also appropriate to report literature reliability values. In all other situations, both literature reliability values and reliability values for data from the current research sample should be reported. Additionally, researchers should remember that single-administration reliability values are not the only methods for addressing reliability and should design their studies to address reliability in a test-retest framework if that is more appropriate for the research context. Regardless of the context, both reliability and validity should be addressed when providing information about data quality to support analyses and interpretations.

Reporting Evidence for Measurement Quality

When reporting evidence for measurement quality, it is important to be very clear about what reliability and validity are properties of. Reliability and validity are not solely properties of the instrument making the measurements, they are properties of the data obtained when using the instrument. Therefore, an instrument itself is never reliable, never valid, and can never be universally validated; those terms should instead be used to describe properties of data obtained from using an instrument in a specific context. In some instances, over time and with a preponderance of reliability and validity evidence, an instrument can become known as a high-quality measurement standard for specific contexts, but that does not mean the instrument is validated for all uses and contexts.

CONCLUSIONS

It can be tempting to interpret the ubiquity of coefficient alpha in both the CER literature and the broader psychology literature[22,23] as evidence that it must be the standard and therefore optimal method for reporting the reliability of psychometric measurements. Unfortunately, this ubiquity results in a vicious cycle where the expectation is that alpha will be reported, even if few reporting or reading the value truly understand what it represents.[9,21] This is especially frustrating given the large

number of studies showing that the mathematical assumptions underlying alpha mean that it frequently is not the most appropriate method for determining single-administration reliability[11,13,15,22,28] and suffers from other flaws such as increasing in value as test length increases.[16,82]

Overreliance on alpha also obscures the fact that alpha is only one of many methods for examining evidence for the reliability of measurements. The analogy between reliability and precision highlights that the ways in which precision is typically expressed in the physical sciences often do not make sense for psychometric measurements because data are rarely obtained from multiple measurements of the same person. However, in situations where two measurements are made using the same instrument, test-retest reliability provides a meaningful summary of the temporal consistency associated with the measurements. Even administering two repeated measures can present logistical and theoretical difficulties for many psychometric variables. The difficulties associated with repeated measures reliability led to the development of alpha and other single-administration reliability coefficients which may be logistically simpler, but come with the tradeoff of stricter underlying assumptions about the measurements.

Of the numerous types of single-administration reliability coefficients that have been developed, alpha is conceptually simple in that it provides information about the proportion of the common construct influencing the true value in an observed measurement relative to the amount of error and also computationally simple[26] in that it can be computed by hand. However, the need for computationally simple reliability coefficients is not nearly as critical as it was in the era before personal computers. As the sophistication of data analysis methodologies have grown, it is worth considering why the choice of a single-administration reliability coefficient has not similarly improved in sophistication. For all its surface simplicity, the utility of alpha is undercut by the rigorous underlying mathematical assumptions, specifically the need for each item to be associated with the common construct to the same degree. Testing for this relation can be done in a factor analysis framework with parallel, tau equivalent, and essentially tau equivalent models. This type of latent variable analysis is becoming more common in CER and opens the door for considering reliability in a factor analysis framework.

McDonald's omega is the most conceptually similar to alpha in that it also provides information about the proportion of common construct value in a measurement relative to the amount of error but, unlike alpha, omega allows each item to be associated with the common construct influencing the true value of each item to a different degree, known as a congeneric model. It is likely that most psychometric instruments best fit a congeneric model, and since omega is mathematically equivalent to alpha if the more restrictive models hold, omega is recommended as a more appropriate single-administration reliability coefficient than alpha for most psychometric measurements. Coefficient H is a factor analysis based approach to reliability that provides a single value to summarize how strongly the items are associated with the common construct, and is applicable to all factor models with good data-model fit, particularly when item responses will be used to calculate weighted scale averages.

Though single-administration reliability coefficients such as alpha, omega, or coefficient H have the advantage of requiring data from only one instrument administration, this does not mean they provide the best information about reliability for every situation. In addition to the types of reliability described in this paper there are other methods for reporting reliability associated with measurements from psychometric instruments including item response theory (IRT) and generalizability theory, the latter of which, also known as G theory, was developed by Cronbach and colleagues.[87] These methods are beginning to gain traction in CER, but further discussion of their benefits and limitations is beyond the scope of this paper, interested readers are encouraged to consult any of the excellent descriptions of these methods.[5,7,26,88–90]

As described in the *Standards*,[3] "there is no single, preferred approach to quantification of reliability/precision. No single index adequately conveys all of the relevant information. No one method of investigation is optimal in all situations" (p. 41). The idea of providing multiple types of evidence to support an argument for measurement quality aligns with the way the CER community has embraced the idea of the multifaceted nature of validity.[1,91] Rather than seeking to provide a single quantification of validity, researchers frequently present many types of validity evidence to provide information about the quality of the data obtained by an instrument. Similarly, reporting of reliability evidence should be considered as one part of a larger set of information, along with validity evidence, that should be reported to provide support for the quality of data obtained from an instrument. Though alpha has

long filled a prominent role in addressing reliability in CER, as best said by Cronbach himself,[26] "I no longer regard the alpha formula as the most appropriate way to examine most data" (p. 403). It is time for CER to consider other options for addressing reliability that are more appropriate for different research contexts and data characteristics.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available on the ACS Publications website at DOI: 10.1021/acs.jchemed.XXXXXXX. [ACS will fill this in.]

R Code for Reliability Calculations (DOCX)

AUTHOR INFORMATION

Corresponding Author

*E-mail: jbarbera@pdx.edu

ACKNOWLEDGMENTS

The authors wish to thank Derek C. Briggs, Program Chair of the School of Education and Director of the Center for Assessment, Design, Research and Evaluation (CADRE) at the University of Colorado at Boulder, Gregory R. Hancock, Program Director of Measurement, Statistics and Evaluation and Director of the Center for Integrated Latent Variable Research (CILVR) at the University of Maryland, and the Portland State University faculty and students attending the Stats Lunch series for their numerous and thoughtful conversations with us about reliability.

REFERENCES

1. Arjoon, J. A.; Xu, X.; Lewis, J. E. Understanding the State of the Art for Measurement in Chemistry Education Research: Examining the Psychometric Evidence. *J. Chem. Educ.* **2013**, *90* (5), 536–545; DOI: 10.1021/ed3002013.
2. Barbera, J.; VandenPlas, J. R. All Assessment Materials Are Not Created Equal: The Myths about Instrument Development, Validity, and Reliability. In *Investigating Classroom Myths through Research on Teaching and Learning*; Bunce, D. M., Ed.; ACS Symposium Series; American Chemical Society: Washington, DC, 2011; Vol. 1074, pp 177–193; DOI: 10.1021/bk-2011-1074.ch011.
3. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. *Standards for Educational & Psychological Testing*; American Educational Research Association: Washington, DC, 2014.

4. Harshman, J.; Yezierski, E. Test–Retest Reliability of the Adaptive Chemistry Assessment Survey for Teachers: Measurement Error and Alternatives to Correlation. *J. Chem. Educ.* **2016**, *93* (2), 239–247; DOI: 10.1021/acs.jchemed.5b00620.
5. Bond, T.; Fox, C. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Mahwah, NJ, 2007.
6. Wright, B. D. Reliability and Separation. *Rasch Meas. Trans.* **1996**, *9* (4), 472.
7. Webb, N. M.; Shavelson, R. J.; Haertel, E. H. Reliability Coefficients and Generalizability Theory. In *Handbook of Statistics*; Rao, C. R., Sinharay, S., Eds.; North Holland: Amsterdam, 2006; pp 81–124; DOI: 10.1016/S0169-7161(06)26004-8.
8. Hallgren, K. A. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor. Quant. Methods Psychol.* **2012**, *8* (1), 23–34.
9. Taber, K. S. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Res. Sci. Educ.* **2017**; DOI: 10.1007/s11165-016-9602-2.
10. Nunnally, J. C.; Bernstein, I. H. *Psychometric Theory*, 3rd ed.; McGraw-Hill: New York, 1994.
11. Dunn, T. J.; Baguley, T.; Brunsden, V. From Alpha to Omega: A Practical Solution to the Pervasive Problem of Internal Consistency Estimation. *Br. J. Psychol.* **2014**, *105* (3), 399–412; DOI: 10.1111/bjop.12046.
12. Haertel, E. H. Reliability. In *Educational Measurement*; Brennan, R. L., Ed.; Praeger Publishers: Westport, CT, 2006; pp 65–110.
13. Cho, E.; Kim, S. Cronbach's Coefficient Alpha: Well Known but Poorly Understood. *Organ. Res. Methods*. **2015**, *18* (2), 207–230; DOI: 10.1177/1094428114555994.
14. Revelle, W.; Zinbarg, R. E. Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. *Psychometrika*. **2009**, *74* (1), 145–154; DOI: 10.1007/s11336-008-9102-z.
15. Yang, Y.; Green, S. B. Coefficient Alpha: A Reliability Coefficient for the 21st Century? *J. Psychoeduc. Assess.* **2011**, *29* (4), 377–392; DOI: 10.1177/0734282911406668.
16. Cortina, J. M. What is Coefficient Alpha? An Examination of Theory and Applications. *J. Appl. Psychol.* **1993**, *78* (1), 98–104; DOI: 10.1037/0021-9010.78.1.98.
17. Green, S. B.; Lissitz, R. W.; Mulaik, S. A. Limitations of Coefficient Alpha as an Index of Test Unidimensionality. *Educ. Psychol. Meas.* **1977**, *37* (4), 827–838; DOI: 10.1177/001316447703700403.
18. Revelle, W. Classical Test Theory and the Measurement of Reliability http://personality-project.org/r/book/# (accessed Jun 2018).
19. Sijtsma, K. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*. **2009**, *74* (1), 107–120; DOI: 10.1007/s11336-008-9101-0.
20. Raker, J. R.; Emenike, M. E.; Holme, T. A. Using Structural Equation Modeling To Understand Chemistry Faculty Familiarity of Assessment Terminology: Results from a National Survey. *J. Chem. Educ.* **2013**, *90* (8), 981–987; DOI: 10.1021/ed300636m.

21. Raker, J. R.; Holme, T. A. Investigating Faculty Familiarity with Assessment Terminology by Applying Cluster Analysis to Interpret Survey Data. *J. Chem. Educ.* **2014**, *91* (8), 1145–1151; DOI: 10.1021/ed500075e.
22. McNeish, D. Thanks Coefficient Alpha, We'll Take It from Here. *Psychol. Methods*. **2017**, 1–22; DOI: 10.1037/met0000144.
23. Crutzen, R.; Peters, G. J. Y. Scale Quality: Alpha is an Inadequate Estimate and Factor-Analytic Evidence Is Needed First of All. *Health Psychol. Rev.* **2015**, *11* (3), 242–247; DOI: 10.1080/17437199.2015.1124240.
24. Green, S. B.; Yang, Y. Commentary on Coefficient Alpha: A Cautionary Tale. *Psychometrika*. **2009**, *74* (1), 121–135; DOI: 10.1007/s11336-008-9098-4.
25. Raykov, T.; Marcoulides, G. A. Thanks Coefficient Alpha, We Still Need You! *Educ. Psychol. Meas.* **2017**, 1–11; DOI: 10.1177/0013164417725127.
26. Cronbach, L. J.; Shavelson, R. J. My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educ. Psychol. Meas.* **2004**, *64* (3), 391–418; DOI: 10.1177/0013164404266386.
27. McDonald, R. P. *Test Theory: A Unified Treatment*; Lawrence Erlbaum Associates: Mahwah, NJ, 1999.
28. Graham, J. M. Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability: What They Are and How to Use Them. *Educ. Psychol. Meas.* **2006**, *66* (6), 930–944; DOI: 10.1177/0013164406288165.
29. Raykov, T.; Marcoulides, G. A. A Direct Latent Variable Modeling Based Method for Point and Interval Estimation of Coefficient Alpha. *Educ. Psychol. Meas.* **2015**, *75* (1), 146–156; DOI: 10.1177/0013164414526039.
30. Wainer, H.; Thissen, D. True Score Theory: The Traditional Method. In *Test Scoring*; Thissen, D., Wainer, H., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, 2001; pp 23–72.
31. Zinbarg, R. E.; Revelle, W.; Yovel, I.; Li, W. Cronbach's α, Revelle's β, and Mcdonald's ω_H: Their Relations with Each Other and Two Alternative Conceptualizations of Reliability. *Psychometrika*. **2005**, *70* (1), 123–133; DOI: 10.1007/s11336-003-0974-7.
32. Hancock, G. R.; An, J. Scale Reliability in Structural Equation Models. *Educ. Meas. Issues Pract.*, in press.
33. Hancock, G. R. Effect Size, Power, and Sample Size Determination for Structured Means Modeling and MIMIC Approaches to between-Groups Hypothesis Testing of Means on a Single Latent Construct. *Psychometrika*. **2001**, *66* (3), 373–388; DOI: 10.1007/BF02294440.
34. Hancock, G. R.; Mueller, R. O. Rethinking Construct Reliability within Latent Variable Systems. In *Structural Equation Modeling: Present and Future: A Festschrift in Honor of Karl Jöreskog*; Scientific Software International: Lincolnwood, IL, 2001; pp 195–216.
35. Byrne, B. M. *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming*, 2nd ed.; Routledge: New York, 2010.

36. Mueller, R. O.; Hancock, G. R. Best Practices in Structural Equation Modeling. In *Best Practices in Quantitative Methods*; Osborne, J. W., Ed.; Sage Publications, Inc.: Thousand Oaks, CA, 2008; pp 488–508; DOI: 10.4135/9781412995627.d38.
37. Kline, R. B. *Principles and Practice of Structural Equation Modeling*, 3rd ed.; The Guilford Press: New York, 2011.
38. Enns, J. T.; Raker, J. R.; Holme, T. A. Validating Chemistry Faculty Members' Self-Reported Familiarity with Assessment Terminology. *J. Chem. Educ.* **2013**, *90* (9), 1130–1136; DOI: 10.1021/ed400094j.
39. *CRC Handbook of Chemistry and Physics*, 98th ed.; Rumble, J. R., Ed.; CRC Press: Boca Raton, FL, 2017.
40. Cronbach, L. J. Test "Reliability": Its Meaning and Determination. *Psychometrika.* **1947**, *12* (1), 1–16; DOI: 10.1007/BF02289289.
41. Schmidt, F. L.; Le, H.; Ilies, R. Beyond Alpha: An Empirical Examination of the Effects of Different Sources of Measurement Error on Reliability Estimates for Measures of Individual-Differences Constructs. *Psychol. Methods.* **2003**, *8* (2), 206–224; DOI: 10.1037/1082-989X.8.2.206.
42. Bandalos, D. L. Methods of Assessing Reliability. In *Measurement Theory and Applications for the Social Sciences*; The Guilford Press: New York, 2018; pp 172–209.
43. Crocker, L.; Algina, J. *Introduction to Classical and Modern Test Theory*; Cengage Learning: Mason, OH, 2006.
44. Spearman, C. Correlation Calculated from Faulty Data. *Br. J. Psychol.* **1910**, *3* (3), 271–295; DOI: 10.1111/j.2044-8295.1910.tb00206.x.
45. Brown, W. Some Experimental Results in the Correlation of Mental Abilitites. *Br. J. Psychol.* **1910**, *3* (3), 296–322; DOI: 10.1111/j.2044-8295.1910.tb00207.x.
46. Feldt, L. S.; Brennan, R. L. Reliability. In *Educational Measurement*; Linn, R. L., Ed.; Macmillan: New York, 1989; pp 105–146.
47. Miller, M. B. Coefficient Alpha: A Basic Introduction From the Perspectives of Classical Test Theory and Structural Equation Modeling. *Struct. Equ. Model. A Multidiscip. J.* **1995**, *2* (3), 255–273; DOI: 10.1080/10705519509540013.
48. Guttman, L. A Basis for Analyzing Test-Retest Reliability. *Psychometrika.* **1945**, *10* (4), 255–282; DOI: 10.1007/BF02288892.
49. Cronbach, L. J. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika.* **1951**, *16* (3), 297–334; DOI: 10.1007/BF02310555.
50. Kuder, G. F.; Richardson, M. W. The Theory of the Estimation of Test Reliability. *Psychometrika.* **1937**, *2* (3), 151–152; DOI: 10.1007/BF02288391.
51. Harshman, J.; Stains, M. A Review and Evaluation of the Internal Structure and Consistency of the Approaches to Teaching Inventory. *Int. J. Sci. Educ.* **2017**, *39* (7), 1–19; DOI:

10.1080/09500693.2017.1310411.

52. Komperda, R.; Hosbein, K. N.; Barbera, J. Evaluation of the Influence of Wording Changes and Course Type on Motivation Instrument Functioning in Chemistry. *Chem. Educ. Res. Pract.* **2018**, *19* (1), 184–198; DOI: 10.1039/C7RP00181A.
53. Liu, Y.; Ferrell, B.; Barbera, J.; Lewis, J. E. Development and Evaluation of a Chemistry-Specific Version of the Academic Motivation Scale (AMS-Chemistry). *Chem. Educ. Res. Pract.* **2017**, *18* (1), 191–213; DOI: 10.1039/C6RP00200E.
54. Xu, X.; Lewis, J. E. Refinement of a Chemistry Attitude Measure for College Students. *J. Chem. Educ.* **2011**, *88* (5), 561–568; DOI: 10.1021/ed900071q.
55. Brandriet, A. R.; Ward, R. M.; Bretz, S. L. Modeling Meaningful Learning in Chemistry Using Structural Equation Modeling. *Chem. Educ. Res. Pract.* **2013**, *14* (4), 421–430; DOI: 10.1039/c3rp00043e.
56. Ferrell, B.; Phillips, M. M.; Barbera, J. Connecting Achievement Motivation to Performance in General Chemistry. *Chem. Educ. Res. Pr.* **2016**, *17* (4), 1054–1066; DOI: 10.1039/C6RP00148C.
57. González, A.; Paoloni, P.-V. Perceived Autonomy-Support, Expectancy, Value, Metacognitive Strategies and Performance in Chemistry: A Structural Equation Model in Undergraduates. *Chem. Educ. Res. Pract.* **2015**, *16* (3), 640–653; DOI: 10.1039/C5RP00058K.
58. Villafañe, S. M.; Xu, X.; Raker, J. R. Self-Efficacy and Academic Performance in First-Semester Organic Chemistry: Testing a Model of Reciprocal Causation. *Chem. Educ. Res. Pr.* **2016**, *17* (4), 973–984; DOI: 10.1039/C6RP00119J.
59. Xu, X.; Villafañe, S. M.; Lewis, J. E. College Students' Attitudes toward Chemistry, Conceptual Knowledge and Achievement: Structural Equation Model Analysis. *Chem. Educ. Res. Pract.* **2013**, *14* (2), 188–200; DOI: 10.1039/c3rp20170h.
60. Bunce, D. M.; Komperda, R.; Schroeder, M. J.; Dillner, D. K.; Lin, S.; Teichert, M. A.; Hartman, J. R. Differential Use of Study Approaches by Students of Different Achievement Levels. *J. Chem. Educ.* **2017**, *94* (10), 1415–1424; DOI: 10.1021/acs.jchemed.7b00202.
61. Salta, K.; Koulougliotis, D. Assessing Motivation to Learn Chemistry: Adaptation and Validation of Science Motivation Questionnaire II with Greek Secondary School Students. *Chem. Educ. Res. Pract.* **2015**, *16* (2), 237–250; DOI: 10.1039/C4RP00196F.
62. Raykov, T. Estimation of Composite Reliability for Congeneric Measures. *Appl. Psychol. Meas.* **1997**, *21* (2), 173–184; DOI: 10.1177/01466216970212006.
63. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing [Computer Software]: Vienna, Austria 2017.
64. Komperda, R. Likert-Type Survey Data Analysis with R and RStudio. In *Computer-Aided Data Analysis in Chemical Education Research (CADACER): Advances and Avenues*; Gupta, P., Ed.; ACS Symposium Series; American Chemical Society: Washington, DC, 2017; pp 91–116; DOI: 10.1021/bk-2017-1260.ch007.

65. Hancock, G. R.; An, J. Scale Reliability in Structural Equation Modeling https://ncme.elevate.commpartners.com/products/digital-module-2-scale-reliability-in-structural-equation-modeling (accessed Jun 2018).
66. Raykov, T. Bias of Coefficient Alpha for Fixed Congeneric Measures with Correlated Errors. *Appl. Psychol. Meas.* **2001**, *25* (1), 69–76; DOI: 10.1177/01466216010251005.
67. Green, S. B.; Hershberger, S. L. Correlated Errors in True Score Models and Their Effect on Coefficient Alpha. *Struct. Equ. Model. A Multidiscip. J.* **2000**, *7* (2), 251–270; DOI: 10.1207/S15328007SEM0702_6.
68. Bandalos, D. L.; Finney, S. J. Factor Analysis: Exploratory and Confirmatory. In *The Reviewer's Guide to Quantitative Methods in the Social Sciences*; Hancock, G. R., Mueller, R. O., Eds.; Routledge: New York, 2010; pp 93–114.
69. Wolf, E. J.; Harrington, K. M.; Clark, S. L.; Miller, M. W. Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educ. Psychol. Meas.* **2015**, *76* (6), 913–934; DOI: 10.1177/0013164413495237.
70. Finney, S. J.; DiStefano, C. Non-Normal and Categorical Data in Structural Equation Modeling. In *Structural Equation Modeling: A Second Course*; Hancock, G. R., Mueller, R. O., Eds.; Quantitative Methods in Education and the Behavioral Sciences: Issues, Research, and Teaching; Information Age Publishing: Charlotte, NC, 2013; pp 439–492.
71. Enders, C. K. Analyzing Structural Equation Models with Missing Data. In *Structural Equation Modeling: A Second Course*; Hancock, G. R., Mueller, R. O., Eds.; Quantitative Methods in Education and the Behavioral Sciences: Issues, Research, and Teaching; Information Age Publishing: Charlotte, NC, 2013; pp 493–519.
72. Raykov, T.; Hancock, G. R. Examining Change in Maximal Reliability for Multiple-Component Measuring Instruments. *Br. J. Math. Stat. Psychol.* **2005**, *58* (1), 65–82; DOI: 10.1348/000711005X38753.
73. Peters, G.-J. Y. *userfriendlyscience: Quantitative Analysis Made Accessible, R package version 0.7.0*; http://userfriendlyscience.com, 2017 (accessed Jun 2018).
74. Kelley, K. *MBESS: An R Package, R package version 4.4.1*; https://www3.nd.edu/~kkelley/site/MBESS.html, 2017 (accessed Jun 2018).
75. semTools Contributors. *semTools: Useful Tools for Structural Equation Modeling, R package version 0.4-14*; https://CRAN.R-project.org/package=semTools, 2016 (accessed Jun 2018).
76. Zhang, Z.; Yuan, K.-H. *coefficientalpha: Robust Coefficient Alpha and Omega with Missing and Non-Normal Data, R package version 0.5*; https://CRAN.R-project.org/package=coefficientalpha, 2015 (accessed Jun 2018).
77. Peters, G.-J. Y. The Alpha and the Omega of Scale Reliability and Validity: Why and How to Abandon Cronbach's Alpha and the Route towards More Comprehensive Assessment of Scale Quality. *Eur. Heal. Psychol.* **2014**, *16* (2), 56–69.

78. Kelley, K.; Pornprasertmanit, S. Confidence Intervals for Population Reliability Coefficients: Evaluation of Methods, Recommendations, and Software for Composite Measures. *Psychol. Methods*. **2016**, *21* (1), 69–92; DOI: 10.1037/a0040086.
79. Gadermann, A. M.; Guhn, M.; Zumbo, B. D. Estimating Ordinal Reliability for Likert-Type and Ordinal Item Response Data: A Conceptual, Empirical, and Practical Guide. *Pract. Assessment, Res. Eval.* **2012**, *17* (3), 1–13.
80. Jackson, P. H.; Agunwamba, C. C. Lower Bounds for the Reliability of the Total Score on a Test Composed of Non-Homogeneous Items: I: Algebraic Lower Bounds. *Psychometrika*. **1977**, *42* (4), 567–578; DOI: 10.1007/BF02295979.
81. Lance, C. E.; Butts, M. M.; Michels, L. C. The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? *Organ. Res. Methods*. **2006**, *9* (2), 202–220; DOI: 10.1177/1094428105284919.
82. Streiner, D. L. Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *J. Pers. Assess.* **2003**, *80* (1), 99–103; DOI: 10.1207/S15327752JPA8001_18.
83. Henson, R. K. Use of Exploratory Factor Analysis in Published Research: Common Errors and Some Comment on Improved Practice. *Educ. Psychol. Meas.* **2006**, *66* (3), 393–416; DOI: 10.1177/0013164405282485.
84. Deng, L.; Chan, W. Testing the Difference Between Reliability Coefficients Alpha and Omega. *Educ. Psychol. Meas.* **2017**, *77* (2), 185–203; DOI: 10.1177/0013164416658325.
85. Tang, H.; Ji, P. Using the Statistical Program R Instead of SPSS To Analyze Data. In *Tools of Chemistry Education Research*; Bunce, D. M., Cole, R. S., Eds.; ACS Symposium Series; American Chemical Society: Washington, DC, 2014; pp 135–151; DOI: 10.1021/bk-2014-1166.ch008.
86. Stanley, L. M.; Edwards, M. C. Reliability and Model Fit. *Educ. Psychol. Meas.* **2016**, *76* (6), 976–985; DOI: 10.1177/0013164416638900.
87. Cronbach, L. J.; Rajaratnam, N.; Gleser, G. C. Theory of Generalizability: A Liberalization of Reliability Theory. *Br. J. Stat. Psychol.* **1963**, *16* (2), 137–163; DOI: 10.1111/j.2044-8317.1963.tb00206.x.
88. Milanzi, E.; Molenberghs, G.; Alonso, A.; Verbeke, G.; De Boeck, P. Reliability Measures in Item Response Theory: Manifest versus Latent Correlation Functions. *Br. J. Math. Stat. Psychol.* **2015**, *68* (1), 43–64; DOI: 10.1111/bmsp.12033.
89. Brennan, R. L. Generalizability Theory. *Educ. Meas. Issues Pract.* **1992**, *11* (4), 27–34; DOI: 10.1111/j.1745-3992.1992.tb00260.x.
90. Briggs, D. C.; Wilson, M. Generalizability in Item Response Modeling. *J. Educ. Meas.* **2007**, *44* (2), 131–155; DOI: 10.1111/j.1745-3984.2007.00031.x.
91. Wren, D.; Barbera, J. Gathering Evidence for Validity during the Design, Development, and Qualitative Evaluation of Thermochemistry Concept Inventory Items. *J. Chem. Educ.* **2013**, *90*

(12), 1590–1601; DOI: 10.1021/ed400384g.