Understanding Coefficient Alpha: Assumptions and Interpretations

Dheeraj Sharma

W.P. No. 2016-03-38
March 2016

INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD-380 015
INDIA

Understanding Coefficient Alpha: Assumptions and Interpretations

Dheeraj Sharma
Indian Institute of Management, Ahmedabad

Abstract

A procedure is developed for determining appropriate levels of scale reliability by examining coefficient alpha in conjunction with the standardized regression coefficient for each variable. Present study aims to examine the effect of addition and deletion of items on coefficient alpha. Also, necessary assumptions for appropriate interpretation of coefficient alpha are examined. Present research suggests that deleting scale items to increase coefficient alpha may result in an under specification of the construct the scale attempts to measure. Furthermore, this research offers prescriptive and descriptive insights for appropriate use of coefficient alpha.

INTRODUCTION

Accuracy in measurement is a cornerstone of knowledge advancement. Advancements in any scientific field are predicated upon defining and measuring objects and concepts. In social sciences, measurement is at the forefront of the advancement of the field, since most of the constructs to be measured are latent variables. Measures reflect these constructs (sometimes closely, sometimes not), but do not directly measure them. The accuracy of these measures is fundamental to the advancement of knowledge. Churchill (1979) identified a process for scale development that has served as the dominant paradigm for the development of marketing measures for over two decades. Fundamental to this topic are the related concepts of reliability and validity.

Reliability has been defined as the extent to which measures are free from error, repeatable, and yield consistent results (Nunnally, 1967 and 1972; Peter, 1979). In other words, reliability envisages the consistency, precision and repeatability of the measures (Kline, 1998). Operationalized, reliability is measured as the proportion of true variance of a construct to the total obtained variance of the data from a scale (Kerlinger, 1986). In marketing, the assessment of reliability has been commonly interpreted as a measure of internal consistency and has been operationalized through coefficient alpha (Cronbach, 1951). Also, “Reliabilities are usually estimated from a single test or, at most, the correlation of one test with alternative form. Consequently, the precision of the reliability estimates needs to be considered” (Nunnally and Bernstein, 1994, p246).

However, the indiscriminant addition or deletion of items in search of a higher coefficient alpha may result in validity problems. Also, extant research demonstrates that coefficient alpha may underestimate reliability (Cronbach, 1951; Osburn, 2000) due to instability arising from various contextual factors. The purpose of the present research is to review existing assumptions that underlie the use and interpretation of coefficient alpha and to provide a framework for its interpretation. Furthermore, we proffer an extension to present procedure to calculate reliability by accounting for the variation associated with the instability arising from repeated measurement.

BACKGROUND

There is less than complete agreement between the definition of reliability and the appropriate process for its assessment. In the marketing literature, coefficient alpha has been the most widely used means of assessing internal consistency (Cortina, 1994, Peterson, 1994). Fundamentally, coefficient alpha assesses the internal consistency of a set of scale items intended to measure a single construct. Most procedures (75%) used for assessing reliability are assessments of internal consistency (Hogan, Benjamin, and Brezinski, 2000).

Marketing measures are frequently described as being "reliable" or "not reliable," though the use of such labels is not so simple. Reliability coefficients measure the reliability of a score derived from a scale used in a particular context, not a definitive assessment of reliability from a scale administered in all varying contexts. The "reliability" of a measure is, in part, a function of the *context* in which it was administered, which explains why a measure may have differing levels of reliability depending upon where and how it was administered. Reliability coefficients are likely to change with each administration of the scale, since the total measure score variance is also likely to change (Scott & Wertheimer 1962; Henson 2001). Such changes may be attributable to any number of factors, including the physical surroundings, in which the scale was administered, the presence of other scale items that may alter the affective state of the respondents, or other situational factors that influence interpretation.

As a result, the assessment of reliability is a process that should be repeated each time a scale is employed. While researchers are encouraged to use measures that have been previously used and found to be reliable, it is important to understand that such findings are context specific. Unless the context for the administration of the scale is a replication of an earlier study, the establishment of scale reliability is a process that should be completed with each administration of a scale.

Nunnally and Bernstein (1994) reliability of a measure should encompass both internal consistency and temporal stability. Plausibly, in addition to internal consistency, a complete assessment of scale reliability should also include the *stability* of a measure over time (Scott & Wertheimer, 1962; Cronbach, 1990; Bohrnstedt, 1993). Stability measures the consistency of response *over a period of time*, provided that the true score of the construct has not changed. A stable measure is one that yields a similar response over time, excluding systematic variation. Most applications of measure stability are drawn from education and psychology (e.g., will an IQ test yield a similar score for subject *i* when administered on two separate occasions.). However, stability is also a relevant consideration in marketing research. For examples, what proportion of the variation in a scale such as SERVQUAL is a function of random fluctuations that are different with each administration of the scale? The purpose of this research is to develop a procedure for determining appropriate levels of scale reliability by examining coefficient alpha in conjunction with the standardized regression coefficient for each variable. The procedure aims to incorporate the variation associated with the inherent instability in any given scale.

Nunnally and Bernstein (1994) promulgate that researchers should aim to reduce the random measurement error to achieve higher "reliability". Also, past research indicates that the stability needs to be incorporated into reliability assessment whenever a measure is to be used more than once with a test group (Thye, 2000). However, if stability is a potential source of random variation, it can be argued that stability should be calculated for even single administrations of a scale to reduce the overall random measurement error. An unstable measure administered once will include errors just as an unstable measure administered in multiple occasions. The failure to provide a second administration of a scale does not make the first administration more reliable. Stability is most often measured through a test-retest procedure, in which the score of one administration of a scale are correlated with the score of a second administration of the scale to the same subjects, with all other elements held constant. In a research setting where only a single administration of a scale is required, test-retest stability can be calculated by drawing a second sample from the same population where the main sample was drawn, thus administering the scale twice within a short time period (to minimize the likelihood of an intervening variable affecting scores.) The administration of the scale should be conducted in a setting parallel to the main study. The degree of correlation of the scale in question between the two administrations to subjects in the second sample would provide a test-retest correlation (i.e., a stability coefficient) for the measure in question.

A reason for the need to calculate a coefficient of stability for marketing measures concerns the nature of error terms. Consider Equation 1, drawn from classical test theory:

(Eq. 1) $X_{\text{true}} = X_{\text{observed}} + E_{\text{internal consistency}}$

When internal consistency measures are used (i.e., coefficient alpha), E is random error that reflects the level of internal inconsistency. This can be shown to be a function of the number of scale items, the degree of inter-item correlation, and the unidimensionality of the scale. When considering a coefficient of *stability*, however, the error term reflects the degree to which changes occur in X_{observed} between observation 1 and 2.

(Eq. 2) $X_{\text{observation-1}} = X_{\text{observation-2}} + E_{\text{stability}}$

Combining both equations to incorporate the change that occurs from observation 1 and 2, it can be shown that

(Eq. 3) $X_{\text{true}} = X_{\text{observed}} + E_{\text{internal consistency}} + E_{\text{stability}}$

(Where $E_{\text{stability}}$ accounts for the changes that occur between observations)

It cannot be assumed that the error term from eq. 1 and eq. 2 are correlated. Indeed, Jenkins and Taber (1977) found that increasing internal consistency has little effect on the stability of a composite score over time, or the composite score's correlation with the true score. This implies that instability having high internal consistency may not translate into high reliability because the instability does not influence internal consistency but may influence the overall reliability.

Further more, both components of reliability can be assumed to be uncorrelated and by default summative in nature; that is, the error detected in each component can be summed to create a more realistic measure of reliability (Henson, 2001). In other words, a measure that has a 0.9 internal consistency coefficient and a 0.8 test-retest reliability coefficient actually has an assumed lower bound of reliability of only 0.7. Plausibly, measurement that does not calculate a test-retest measure of stability makes the implicit assumption that the measure is perfectly stable – that instability does not contribute to error variance. As a result, any variance due to instability may be incorrectly modeled as true score variance and provide a potential source of bias to the score.

The need to obtain a second, parallel sample from the same population has likely deterred many researchers from attempting to determine a coefficient of stability through test-retest reliability. Calculation of a coefficient of stability is a judgment call, based on the importance of the accuracy of measurement (Nunnally and Bernstein, 1994). For instance, the calculation of stability may be particularly more important when measuring *traits*, compared to *states* that may vary contextually (Nunnally and Bernstein, 1994). A researcher may wish to consider the merits of drawing a slightly smaller initial sample and then drawing a second sample for use in calculating the coefficient of stability, where he/she suspects there may be an issue of stability or when the accuracy of measurement is particularly critical. The tradeoff of this calculation with a loss of power can only be made by comparing the loss of statistical power from a smaller sample with the benefit of an assessment of measure stability.

BASIC ASSUMPTIONS OF COEFFICIENT ALPHA

Coefficient alpha is calculated through Equation 4.

$$\text{(Eq. 4)} \qquad \alpha = \frac{[N^2 \bullet M(COV)]}{(\text{SUM of VAR/COV})}$$

N represents the number of scale items, M(COV) is the mean inter-item covariance, and SUM of VAR/COV is the sum of all of the elements in the variance/covariance matrix (Cortina, 1993). A common way of interpreting coefficient alpha is to use Nunnally's 1972 recommendation that alpha levels should be 0.7 or higher (Peterson 1994). However, this approach may be very simplistic and may preclude the understanding of the basic assumptions of coefficient alpha and its proper interpretation. Following are the basic assumptions of coefficient alpha:

Assumption 1: Coefficient alpha assumes the scale is continuous. First, it is important to understand the coefficient alpha is intended for use with continuous scales. If the goal is to assess the reliability of a dichotomous measure, the appropriate measure of internal consistency would be a KR-20 test (Kuder & Richardson 1937). Both coefficient alpha and KR-20 can be interpreted as the average of all possible split-half corrected reliability estimates for a given instrument. This is clearly preferable to a single split-half procedure, in which the chances exist for an outlying reliability coefficient, based on poor composition of two "parallel" test halves.

Assumption 2: Coefficient alpha assumes that scale items are tau-equivalent. Coefficient alpha has its roots in classical test (CT) theory. CT theory recommends that a coefficient of equivalence (strict or random) be calculated for a measure (Bohrnstedt, 1993). Generally speaking, there is little application for a coefficient of strict equivalence, since such a calculation would require two perfectly parallel versions of the same test. Coefficient alpha and KR-20 are both coefficients of random equivalence and require that measures be tau-equivalent. (Tau-equivalence means that the scale items vary from the true score by only a constant.). If *n* items are all essentially tau-equivalent, then coefficient alpha is exactly equal to the reliability of the measure (Lord & Novick, 1968). The degree to which this tau-equivalent assumption is not met influences the accuracy of coefficient alpha. More specifically, when measures are not all tau-equivalent, coefficient alpha is under-reported and becomes a lower bound of actual reliability (i.e., reliability is under-reported for the measure) (Novick & Lewis, 1967).

The tau-equivalent assumption is particularly important for scales with three or fewer items. In such cases, reliability was under-reported by .03 to .11, depending upon the degree of violation (Raykov, 1997). For larger scales (four or more items), having a single item that is not tau-equivalent produces only a moderate lowering of coefficient alpha (0.03 or less). This helps to clarify earlier results that scales with three or fewer items reported significantly lower reliability scores than those with more than four items (Churchill & Peter, 1984). Ostensibly, the coefficient alpha levels for scales with three or fewer items need to be examined relative to possible underreporting of their reliability levels surfacing. In particular, alternative measures for assessing internal consistency should be considered under such circumstances, including factor analytic methods (Miller, 1995)

<u>Assumption 3: Coefficient alpha assumes the scale is unidimensional.</u> Coefficient alpha assumes that the entire variance associated with the measure is drawn from a single latent construct. Adding additional items to a scale in order to increase Cronbach's alpha may increase reliability, but at the cost of validity.

Unidimensionality cannot be inferred from a relatively high coefficient alpha. Gilner (2001) has pointed out that a coefficient alpha of 0.9 for a scale with a large number of items (e.g., 30 items) indicates neither a high inter-item correlations nor a single underlying dimension; because the coefficient alpha numerator specifically contains the number of scale items in it, and it is not a test for unidimensionality. Cortina (1993) demonstrated that a scale with more than 14 items may have a coefficient alpha of 0.70, even if it is measuring two dimensions and the item inter-correlations average only 0.3 (Cortina, 1993). As the number of scale items increases, there exists the possibility for a high coefficient alpha even though the items are assessing more than a single dimension.

This raises the issue of whether reliability is a precursor to the assessment of construct validity, or whether reliability is a part of the validity assessment process. It has been often stated that a measure has to be reliable in order to be valid, but that the inverse is not true. Churchill’s (1979) paradigm for scale development implicitly states that measures need to be determined reliable before a claim of validity can be assessed. However, if an assessment of unidimensionality is needed to determine scale reliability, and an assessment of unidimensionality is included in the construct validation process, then it may be necessary to think of the assessment of reliability as a part of the construct validation process and not as a precursor to that process.

UNDERSTANDING AND INTERPRETING COEFFICIENT ALPHA

In marketing, the two most common means of interpreting coefficient alpha are a) how the measure’s reliability coefficient compares to similar measures (Peterson, 1994), and b) the context of the type of research in which the measure is to be used (Nunnally 1972). However, marketing research widely used rules of thumb as a criterion for assessment of reliability. Nunnally (1967) set the standards for reliability coefficients used in *basic* research at 0.6, then later increased this to 0.7 (1972). The effect of Nunnally’s criteria for assessing reliability in marketing has been considerable. Peterson (1994) has noted that the standards were specifically cited over 50 times in a twelve-year period in the *Journal of Marketing Research* alone. Further evidence of the importance of these standards can be found in an examination of reported coefficient alpha levels – 75% of the alpha coefficients in published journals met or exceeded Nunnally’s 0.70 criteria (Peterson, 1994).

EFFECTS OF LOW LEVELS OF COEFFICIENT ALPHA

The establishment of these levels is not definitive, however. Nunnally (1972) indicates that during the theory building/construct definition processes of research, the effect of increasing reliability coefficients from 0.7 to higher levels (above 0.8) is not likely to be worth the additional effort, given the modest effect on correlations presented by measurement errors when reliabilities exceed 0.7. As evidence, Nunnally (1972) provided a process for examining the effect of low reliability on correlations, using the following formula (Eq. 5) (Nunnally, 1972, p. 238)

(Eq. 5) $r_{\text{expected - xy}} = r_{xy} \bullet \sqrt{((r'_{xx} \bullet r'_{yy})/(r_{xx} \bullet r_{yy}))}$

where $r_{\text{expected - xy}}$ is the expected correlation between variables X and Y, r_{xy} is the observed correlation between variables X and Y,) r_{xx} is the actual reliability coefficient for scale X, r_{yy} is the actual reliability coefficient for scale Y, and r'_{xx} is the changed reliability for variable X, and r'_{yy} is the changed reliability for variable Y. For example, if the observed correlation between X and Y is $r = 0.25$, the observed α for X is 0.7 and the α for Y is 0.65, and the changed reliability coefficients for both measures are 0.85, the corrected correlation is

$$r_{\text{exp-xy}} = (0.25) \bullet \sqrt{((0.85 \bullet 0.85)/(0.7 \bullet 0.65))} = 0.315$$

Given a reasonable sample size, the argument is that such a modest change in a correlation would not likely change the resulting p-value for a correlation coefficient. This argument has been shown to be partial (Cortina, 1993; Raykov, 1997; Thye, 2000) Lower levels of reliability reflect a greater level of measurement error, which can weaken the correlation between the variable of interest and other variables (Cortina 1993; Raykov 1997; Thye, 2000). This can result in a Type II error, the failure to detect a significant relationship between two variables when such a relationship does, in reality, exist. This occurs through a lowering of power (Cohen, 1969).

However, the failure to control random errors (which leads to lower levels of reliability) can also lead to a Type I error (Thye, 2000). For example, assume that an experimental group and a control group are both completing a scale that has a modest level of reliability (0.70), and that the true scores are the same for both groups. The low reliability scale is administered, means are calculated for the experimental group and the control group, and the difference between the two means is found to be statistically significant. Is the difference a function of the treatment, or a function of the low reliability of the measurement instrument? The use of a measure with a lower level of reliability may result in the detection of a difference where none exists – a Type I error. This is particularly true when a large sample is used and relatively modest changes in effect sizes are statistically significant. As a result, absolute standards for interpreting the appropriateness of coefficient alpha are only useful as general guidelines.

INCREASING COEFFICIENT ALPHA

Previous research (Churchill & Peter, 1984; Peterson, 1994) reported a weak relationship between the number of scale items and coefficient alpha used in marketing research, while others have noted that reliabilities will increase by increasing the number of scale items (see Nunnally, 1972; Jenkins & Taber, 1977; Cortina, 1993; Raykov, 1997; Keller & Dansereau, 2001). Since the number of scale items is specifically modeled in the numerator of coefficient alpha, reliability levels will increase as the number of scale items increase, *assuming everything else is constant,* and herein lies the reason for the discrepancy between previous research findings and theory. The specific effect of increasing the number of scale items on coefficient alpha is shown in Eq. 6 (Nunnally, 1972):

(Eq. 6) $$r_{kk} = (k \bullet r_{xx})/(1 + (k - 1) \bullet r_{xx})$$

where k is the factor of increase for the scale items and r_{xx} is the reliability coefficient. If the number of scale items was increased by 50% (e.g., from 4 to 6) and the reliability coefficient for the four-item scale was 0.7, then making the scale a six-item measure would increase the reliability coefficient to 0.778

$$\mathrm{r}_{kk} = (1.5 \bullet 0.7)/(1 + (1.5 - 1) \bullet 0.7) = 0.778$$

However, the above procedure renders at least two assumptions that may not be justified. First, the equation assumes that even with adding two scale items, the average correlations among scale items will remain constant. Second, the equation also assumes that the two added scale items will represent the same single dimension as the first four items. For the added scale items to increase levels of coefficient alpha, it is important that both assumptions are met.

Another, common solution to increasing scale reliability is through the elimination of items that have low item-to-total correlations. In cases where a scale item does not load onto the same single dimension with other scale items, such a decision is likely appropriate. However, while the deletion of items with low item-to-total correlations can increase scale reliability, they can also result in an under-identification of the construct (Raykov, 1997; Smith, 1999). In search of higher reliabilities, scale developers have been encouraged to increase the number of individual scale items and to make these items as similar as possible to existing items – this will result in a unidimensional structure and high item-to-total correlations. Combined with an increased number of scale items, the net effect is an increase in scale reliability. However, it can be argued that deleting all items with lower item-to-total correlations and adding scale items that are similar to those already in a measure serves to increase reliability at the risk of construct validity. Removing all scale items with low item-to-total correlations is likely to eliminate some true score variance, along with error variance (Raykov, 1997).

Adding and deleting items to increase coefficient alpha should not be undertaken without due diligence. By adding scale items in order to increase scale reliability, we may inadvertently change the definition of a construct by over-sampling from a portion of it (Keller & Dansereau, 2001). Adding items to a scale in order to improve the reliability of the measure can result in a reduction of predictive validity. At the same time, deleting items to improve reliability levels solely on the basis of the SPSS's "alpha if item deleted" is also inappropriate. Much of the true score variance may also be eliminated in doing so.

The current scale development process may not prevent this type of error. It is possible that this may be prevented by the face validity assessment of the items by experts. But normally, the experts assess the representativeness of the items for the domain, not the issue of coverage for the entire domain. In addition, this assessment is performed prior to data collection, when alpha calculations are not available.

The notion of coverage of the domain is not really part of the current analysis of measures used in marketing. On the other hand, Item response theory (IRT) has long considered the range of the latent trait important. Like current marketing analysis, individuals are scored on the construct of interest. In addition, with IRT, measurement items are simultaneously scaled on the same construct (Lord & Novick, 1968). In other words, the items constituting the measurement are assessed and placed along the continuum of the domain of the construct. This provides evidence of coverage of the domain or bandwidth in IRT terms.

This suggests that domains should be measured with scales using items that tap the entire range of the domain or the entire bandwidth. This may lead to a lower level of internal consistency, which is good in this case. The desire for a higher alpha should not be the single focus, for it might lead to the creation of validity problems.

Without concern for the entire domain of the construct, the result may be a measure that demonstrates high item-to-total correlations, but one in which the breadth of the construct is under-represented. The effect is a threat to construct validity. Such under-specified measures are likely to demonstrate relationships to other variables in a model that differ from their true relationships with those constructs.

METHODOLOGY

An Example Using Multiple Linear Regression

As previously defined, the reliability of a measure is operationalized as the proportion of true score variance to the total obtained variance (Kerlinger, 1986). That is, a measure that possesses a reliability coefficient of 0.80 is assumed to have 20% of its variance attributable to random measurement factors.

To determine the effects of varying degrees of coefficient alpha on the amount of random measurement error, a series of simulated regression analyses was conducted. In each analysis, a multiple regression equation featuring three independent variables was created of the form $Y = a + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ where β_i represents standardized regression coefficient for variable X_i. The variables (X_i) are scale scores created by averaging the multi-item scales for each variable.

Standardized regression coefficients are particularly appropriate for such analysis. By comparing sizes of the coefficient relative to one another, the standardized regression coefficients can indicate the importance each independent variable to the prediction of the dependent variable. Also, note that in the present example, there is no error term. Error is modeled at 0, since the intent is to identify the amount of random measurement error independent of other error (e.g., sampling). Standardized regression coefficients were generated using random numbers, with the size of the coefficients constrained so that the sum of the three standardized regression would always equal 1.0.

Reliability levels were fixed for 512 equations, covering all possible combinations for three measures with alpha coefficients of 0.95, 0.90, 0.85, 0.80, 0.75, 0.70, 0.65 and 0.60. Reliability levels below 0.60 were not modeled in the analysis, since those have been judged for 30+ years to be insufficient for even basic research (Nunnally 1972). For example, equation #1 featured coefficient alpha levels of 0.95, 0.95, and 0.95 for all three independent variables. The coefficient alpha levels used in Equation #2 for the three variables were 0.95, 0.95 and 0.90, respectively, while Equation #3 coefficient alpha levels were 0.95, 0.95, and 0.85. This was repeated for all 512 possible combinations of reliability coefficients for the three variables.

To determine the amount of error generated due to random measurement error, an equation for calculating an estimate of random measurement error is needed. If $r_{tt\text{-}i}$ (coefficient alpha) is interpreted as a measure of the reliability of *i*, then 1 - $r_{tt\text{-}i}$ estimates the random measurement error associated with measure *i*. The relative amount of random measurement error contributed by each independent variable towards the total measurement error is a function of the importance of each independent variable in predicting/explaining the dependent variable. One means of assessing the relative importance of each independent variable is an examination of standardized regression coefficient, β_i for each independent variable. Since β_i can be interpreted as a measure of the relative importance of an independent variable to the dependent variable, the product of β_i and 1 - $r_{tt\text{-}I}$ is an estimate of the amount of random measurement contributed by variable *i*. By summing this across the independent variables in the equation, an estimate of random measurement error is derived. This is shown in Eq. 7:

$$\text{Total random measurement error} = \Sigma((1-r_{tt\text{-}i})(\beta_i)) \qquad \text{(Eq. 7)}$$

It should be noted that Eq. 7 would yield an estimated upper bound of random measurement error. Eq. 7 assumes that the error terms for each independent variable are uncorrelated. The degree to which these error terms are correlated would lower the calculation of random measurement error.

Since most marketing measures are used in conjunction with other measures to explain the relationship between two or more independent variables with a dependent variable, the effects of varying level of reliability coefficients to total random measurement error was also examined. A sample of the results is shown in Table 1.
<Insert Table 1 here>

It is important to recognize that when several scales are being used in the same analysis and scale reliabilities are dissimilar (i.e., some are high (0.9 or greater) and some are low (0.7 or lower), the influence of coefficient alpha on total measurement error is a function of a) the number of variables present and b) the importance of each variable as a predictor of the dependent variable (i.e., the size of the standardized regression coefficient). As more variables are added to the equation, the need for each measure to be reliable increases, if the researcher is to have confidence in the resulting levels of the dependent variable. The relative size of each standardized regression coefficient is also an influence. As an independent variable increases in importance to an overall equation (as evidenced by a larger standardized regression coefficient), so does the importance of reliability for that measure. As a result, it is important that when examining the appropriateness of a level of reliability, the *specific context* in which it is being applied must be considered.

To illustrate, consider $Y = a + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + e$, under circumstances where β_1 is considerably larger than β_2 or β_3 (for example, if β_2 is 0.6 while β_2 and β_3 are

0.1 each) and coefficient alpha for β_1 is relatively low (e.g., 0.7) compared to the alpha coefficient for β_2 and β_3 (e.g., both at 0.9). β_1 contributes 75% of the explained variance in Y (0.6/0.6 + 0.1 + 0.1), while β_2 and β_3 contribute only 12.5% each. Yet β_1 contributes a disproportionately greater amount of random measurement error. From Eq. 7 we can see that the total random measurement error is .20 [$\Sigma((1\text{-}r_{tt})(\beta_i))$ = (1-0.7)(.6) + (1-0.9)(.1) + (1-0.9)(.1) = .20]. By modifying Eq. 7 to convert the β_i into a proportion of random measurement error (i.e., $(1\text{-}r_{tt})(\beta_i)/\Sigma((1\text{-}r_{tt})(\beta_i))$, the proportion due to β_1 is 90% (.18/.20=.90). So while β_1 explains 75% of the variance in the dependent variable, it provides 90% of the random measurement error.

What if a lower level of reliability (.7) was reported for the measure associated with β_2, while β_1 and β_3 had reliability coefficients of .9? Keeping all other numbers the same, the total random measurement error is now .10 [(1-0.9)(.6) + (1-0.7)(.1) + (1-0.9)(.1)]. While β_2 still contributes 12.5% of the explained variance, it now provides 30% of the random measurement error (.03/.10=.30) as opposed to 5% in the previous example. In this case, the lower level of reliability is not near as damaging to the total random measurement error due to the lower level of importance of β_2. In addition, β_1 contributes 60% of the random measurement error, but with a higher level of reliability and thus total random measurement error is lower.

In short, the size of a standardized regression coefficient underscores the relative importance of measure reliability to determining total random error present. Especially considering the disproportionate effects on variance explained and random error, with the more important, less reliable variable contributing much more to random error than variance explained. The more important the independent variable, as evidenced by its standardized regression coefficient, the more important it is for a measure to possess a higher degree of reliability to avoid such a contribution to random measurement error.

RECOMMENDATIONS/CONCLUSIONS

For over fifty years, the assessment of scale reliability has been a part of the construct validation process. During this time, statements like *reliability coefficients* *need to exceed 0.70 for basic research* and *reliability is a prerequisite for validity* have been imprecisely used in presenting results of marketing research. The present research seeks to bring issues of reliability back to the forefront.

At a basic level, a reliable measure is one that is both internally consistent and predictable in repeated administrations of the scale. Although, coefficient alpha is a popular and fairly robust means of assessing the *internal consistency* of a measure (Iacobucci & Duhachek, 2003), but researchers need to also consider an assessment of the *stability* of a measure. Such an assessment is particularly important, since both assessments are likely to identify independent sources of random error that are uncorrelated with one another. We concur with previous research that suggests that assessment of stability of a measure may be very important in case of trait measures that are intended to be reliable (Nunnally and Bernstein, 1994).

Researchers need to understand that coefficient alpha needs to be interpreted relative to the context in which the scale is used. First, coefficient alpha is appropriate only when scale items are continuous. Second, coefficient alpha assumes that a measure must be unidimensional. Third, coefficient alpha assumes that scale items are tau-equivalent. Violation of these conditions will bias coefficient alpha levels. In such cases, a researcher is advised to consider alternative means of assessing scale reliability.

In addition, the authors call for scale users to assess the coverage of the domain of the construct. Under-representativeness of the domain may result from adding or deleting items without a concern for domain coverage. Currently in marketing, assessment of this would follow in the same manner as face validity. Experts would be provided with a definition of the domain and charged with assessing the ability of the set of items to adequately tap the entire domain. While IRT programs are available (see Roberts & Laughlin, 1996 or Andrich, 1996) for examples), they have not been used widely in marketing. As a result, a first step would be for marketers to evaluate domain coverage with experts, with a goal of moving toward the use of IRT techniques to quantitatively assess bandwidth.

The present research shows that determining the appropriate level for coefficient alpha is a function of the context in which the measure is used. Previous research (e.g., Nunnally, 1972) has identified that alpha coefficients need to be interpreted in terms of the *purpose* of the research setting. The present research extends this by demonstrating that alpha coefficients also need to be interpreted in terms of the importance of the measure to the analysis. Measures that contribute a greater proportion of the unique variance in the dependent variable need to have higher levels of reliability. In this study, regression was used to examine the standardized regression coefficients in conjunction with coefficient alpha to determine the error variance contributed by each measure in the model.

This type of analysis may be particularly important when one collects data and finds at least one scale has a reliability that is potentially troublesome. Upon further analysis, the researcher may find that the importance of the variable is low, in which case the lower level of reliability is not as troublesome. As a general guideline, researchers need to be alert to instances where the standardized regression coefficient of an independent variable is two or more times the size of any other independent variable in the same analysis. Such a circumstance indicates a need to re-examine the level of reliability for the measure of an independent variable so important in the explanation of the dependent variable. A level of reliability higher than that prescribed by Nunnally (is necessary for such a measure, since it contributes a greater proportion of the total random error for the equation. Further research to determine what combinations of alpha and variable importance are acceptable and what combinations are troublesome would be a next step.

Determining an appropriate level of coefficient alpha for a given measure is a complex issue (Iacobucci & Duhachek, 2003). It has been long accepted in consumer research that it is important to consider whether the scale will be used in basic or applied research. Also, the education and psychology fields are replete with examples of the need for high levels of reliability (Aiken, 2002, Anastasi & Urbina, 1996). In consumer research, most scales are used in conjunction with other measures to create models to predict or explain something. The use of multiple measures of different constructs requires progressively higher levels of reliability, since the likelihood of making a Type II (or in some cases, a Type I error) increases with the random error associated with each additional measure included in a model. Hence, strengthening the measures of reliability may be critical to attaining meaningful and precise results.

References

Aiken, L.R. (2002). *Psychological testing and assessment*. (11th ed.). Boston, Allyn & Bacon

Anastasi, M.J., and Urbina, S. (1996). *Psychological testing* (7th ed.) New York: Prentice-Hall

Andrich, D. (1996). A Hyperbolic Cosine Latent Trait Model for Unfolding Polytomous Responses: Reconciling Thurston and Likert Methodologies. *British Journal of Mathematical and Statistical Psychology*, 49, 347-365.

Bohrnstedt, G. (1993). *Classical Measurement Theory: Its Utility and Limitations for Attitude Research*. in New Directions in Attitude Measurement, Dagmar Krebs and Peter Schmidt, eds. Berlin: Walter de Gruyter & Co.

Churchill, G. A., Jr. (1979). A Paradigm for Developing Better Measures of Marketing Constructs. *Journal of Marketing Research*, 16 (February), 64-83.

Churchill, G.A., and Peter, P.J. (1984). Research Design Effects on the Reliability of Rating Scales: A Meta-Analysis. *Journal of Marketing Research,* 21 (November), 360-375.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cortina, J. M. (1993). What Is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*, 78 (1) 98-104.

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16 (September), 297-334

Groth-Marnat, G. (2003*). Handbook of Psychological Assessment*. Hoboken, New Jersey, John Wiley & Sons

Groth-Marnat, G. (1990). *Essentials of Psychological Testing*, (5th Ed.). New York: Harper and Row

Gilner, J. A., Morgan, G. A., and Harmon, R. J. (2001). Measurement Reliability. *Journal of the American Academy of Child and Adolescent Psychiatry*, 40 (April), 486-488.

Henson, R. K. (2001). Understanding Internal Consistency Reliability Estimates: A Conceptual Primer on Coefficient Alpha. *Measurement and Evaluation in Counseling and Development*, 34 (October), 177-189.

Hogan, T. P., Benjamin, A., and Brezinski, K. L. (2000). Reliability Methods: A Note of the Frequency of Use of Various Types. *Educational and Psychological Measurement*, 60, 523-531

Iacobucci, D. and Duhachek, A. (2003). Advancing Alpha: Measuring reliability with confidence. *Journal of Consumer Psychology*. 13(4), 478-487

Jenkins, G. D., and Taber, T. D., (1977). A Monte Carlo Study of Factors Affecting Three Indices of Composite Scale Reliability. *Journal of Applied Psychology*, 62 (4), 392-398.

Keller, T., and Fred D., (2001). The Effect of Adding Items to Scales: An Illustrative Case of LMX. *Organizational Research Methods*, 4 (April), 131-143.

Kerlinger, F. N., (1986), *Foundations of Behavioral Research*, (3rd Ed.), New York: Holt Rinehart, and Winston.

Kline, R.B. (1998), *Principles and Practice of Structural Equation Modeling*, New York: The Guilford Press

Kuder, F., and Richardson, M. W., (1937). The Theory of the Estimation of Test Reliability. *Psychometrika*, 2, 135-138.

Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*, Reading, MA: Addison Wesley.

Miller, M. B. (1995). Coefficient Alpha: A Basic Introduction from the Perspective of Classical Test Theory and Structural Equation Modeling. *Structural Equation Modeling*, 2, 25-273.

Novick, M. R. and Charles L. (1967). Coefficient Alpha and the Reliability of Composite Measurements. *Psychometrika*, 32, 1-13.

Nunnally, J.C. (1967). *Psychometric Theory*, (1st Ed.). New York: McGraw-Hill.

Nunnally, J.C. (1972), *Psychometric Theory*, (2nd Ed.). New York: McGraw-Hill.

Nunnally, J.C. and Bernstein, I. H. (1994), *Psychometric Theory*, (3rd Ed.). New York: McGraw-Hill.

Peter, J. P., (1979). Reliability: A Review of Psychometric Basics and Recent Marketing Practices. *Journal of Marketing Research*, 16 (February), 6-17.

Raykov, T., (1997). Scale Reliability, Cronbach's Coefficient Alpha, and Violations of Essential Tau-Equivalence with Fixed Congeneric Components. *Multivariate Behavioral Research*. 32 (4), 329-353.

Roberts, J. S. and Laughlin, J. E., (1996). A Unidimensional Item Response Model for Unfolding Responses from a Graded Disagree-Agree Response Scale. *Applied Psychological Measurement.* 20, 231-255.

Scott, W. A. and Wertheimer, M. (1962). *Introduction to Psychological Research*, New York: John Wiley and Sons, Inc.

Smith, Anne M. (1999). Some Problems When Adopting Churchill's Paradigm for the Development of Service Quality Measurement Scales. *Journal of Business Research*, 46 (October), 109-120

Thye, S. R. (2000). Reliability in Experimental Sociology. *Social Forces*, 78 (June), 1277-1309.

Table 1
Error Ranges by Level of Reliability in a Regression Equation with Three Independent Variables

Coefficient alpha for three variables			***Lowest* possible bounded random measurement error**	***Highest* possible bounded random measurement error**	**Mean[1] measurement error**
α_1	α_2	α_3			
.95	.95	.90	.05	.10	.073
.95	.80	.80	.05	.20	.153
.95	.70	.70	.05	.30	.220
.95	.60	.60	.05	.40	.287
.90	.80	.80	.10	.20	.180
.90	.70	.70	.10	.30	.220
.90	.60	.60	.10	.40	.343

[1] Three separate regression equations were run, each with different, randomly generated regression coefficients. The mean measurement error here is the mean of these three error calculations, based on $\Sigma((1-r_{tt-i})(\beta_i))$