**Educational and**
**Psychological Measurement**
Volume 66 Number 6
December 2006 930-944
© 2006 Sage Publications
10.1177/0013164406288165
http://epm.sagepub.com
hosted at
http://online.sagepub.com

# Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability

## What They Are and How to Use Them

James M. Graham
*Western Washington University*

Coefficient alpha, the most commonly used estimate of internal consistency, is often considered a lower bound estimate of reliability, though the extent of its underestimation is not typically known. Many researchers are unaware that coefficient alpha is based on the essentially tau-equivalent measurement model. It is the violation of the assumptions required by this measurement model that are often responsible for coefficient alpha's underestimation of reliability. This article presents a hierarchy of measurement models that can be used to estimate reliability and illustrates a procedure by which structural equation modeling can be used to test the fit of these models to a set of data. Test and data characteristics that can influence the extent to which the assumption of tau-equivalence is violated are discussed. Both heuristic and applied examples are used to augment the discussion.

***Keywords:*** *reliability; structural equation modeling; congeneric; tau-equivalent*

A number of studies have shown that ignorance regarding fundamental measurement issues has reached an endemic level (Vacha-Haase, Kogan, & Thompson, 2000; Whittington, 1998). Although many doctoral programs include exposure to statistics and research design, measurement issues are often ignored in education and psychology programs. As a result, issues such as reliability are often misconstrued (Aiken, West, Sechrest, & Reno, 1990; Pedhazur & Schmelkin, 1991). Although some progress has been made in educating researchers about reliability, such as the dissemination of the fact that reliability is an artifact of the sample, not the test (Thompson & Vacha-Haase, 2000), many researchers still lack the basic knowledge necessary to accurately estimate reliability.

The most commonly used measure of internal consistency, coefficient alpha, is based on the essentially tau-equivalent measurement model, a measurement model that requires a number of assumptions to be met for the estimate to accurately reflect

**Author's Note:** Please address correspondence to James M. Graham, Department of Psychology, Western Washington University, 516 High Street, Bellingham, WA 98225-9089; e-mail: jim.graham@wwu.edu.

the data's true reliability (Raykov, 1997a). Violation of these assumptions causes coefficient alpha to underestimate the true reliability of the data (Miller, 1995). In fact, there are a number of other measurement models, "parallel," "tau-equivalent," and "congeneric" (Feldt & Brennan, 1989, pp. 110-111; Lord & Novick, 1968, pp. 47-50), that can be used to estimate reliability.

This article describes these measurement models as they apply to classical test theory and reliability. The use of structural equation modeling (SEM) path diagrams to evaluate the fit of these models is described. A procedure for obtaining an accurate estimate of reliability based upon these findings is outlined. Finally, test and data characteristics that influence the extent of coefficient alpha's underestimation of reliability are presented. Whereas previous work has explored the theoretical underpinnings of these concepts in greater depth and complexity (Miller, 1995; Raykov, 1997a, 1997b), it is my intention to describe these concepts in a more accessible format with ample use of both heuristic and applied examples.

## Reliability in Structural Equation Modeling

Classical test theory (CTT) is based on the premise that the variance in observed scores ($X$) is due in part to true differences in the latent trait being measured ($T$) and in part to error ($E$). This can be represented in the equation $X = T + E$. This basic equation can be represented in an SEM path diagram in which one measured variable, $X$, and two latent variables, $T$ and $E$, are shown in relation to one another. Unidirectional paths (with path coefficients set to 1) from the latent variables to the measured variable indicate that measured scores are an additive combination of the true and error scores. Additionally, the latent variables are uncorrelated with one another.

In CTT, a reliability coefficient ($\rho_{xx}$) is the proportion of observed score ($X$) variance accounted for by the true score ($T$) variance, as shown by the following equation (Miller, 1995):

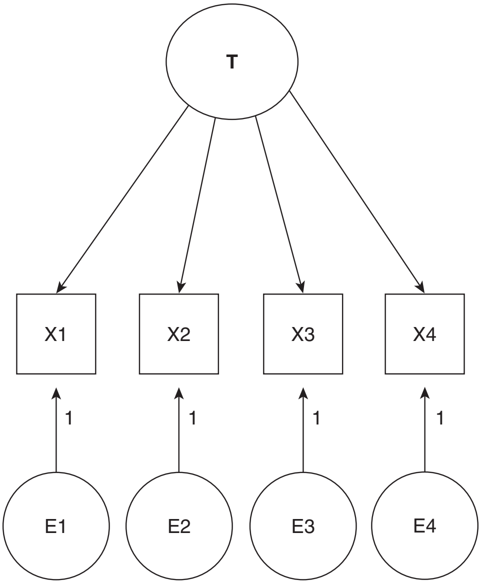
$$\rho_{xx} = \mathrm{Var}(T)/\mathrm{Var}(X). \tag{1}$$

The present discussion focuses on one type of reliability, internal consistency, as opposed to other types of reliability, such as test-retest, alternate forms, or interrater.

The SEM model described above is not identified. As no parameters are given for the partitioning of the observed score variance into the two latent variables, there are an infinite number of ways in which the variance of $X$ can be divided into $T$ and $E$. Therefore, it is impossible to estimate the reliability of $X$ unless the test consists of more than one item. In multiple-item tests, each item has its own true and error scores; again, without further specifications, there are still an infinite numbers of ways to partition the variance of items. To estimate the reliability of these items, we must identify the model by making further assumptions.

Estimates of reliability within CTT assume that all observed variables measure a single latent true variable. Many researchers erroneously believe that reliability provides a

**Figure 1**
**Structural Equation Modeling (SEM) Path Diagram of Unidimensional Composite True Variable**



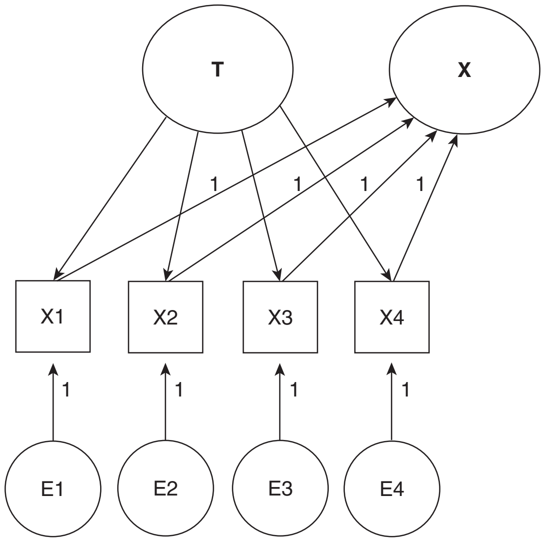
measure of test unidimensionality. In actuality, reliability assumes that unidimensionality exists (Miller, 1995). Failure to meet the assumption of unidimensionality will result in an inaccurate and often misleading estimate of reliability.

The assumption of unidimensionality for a four-item test is represented as an SEM path diagram in Figure 1. Here the variables are shown with subscripts in the text, for example, $X_1$, but without subscripts in the figures, for example, *X*1. These refer to the same variables. The model in Figure 1 shows a single latent variable (*T*) as being responsible for part of the observed score variances of each individual test item ($X_1$, $X_2$, etc.). Additionally, each item has a unique error term associated with that item. In the model shown in Figure 2, there are no numbers on the paths from the latent true variable to the individual item observed variables. The relationship between the composite true score and item true scores can be defined a number of ways and will be dealt with at a later point.

## Estimating Reliability Within SEM

As previously noted, reliability is the proportion of true to observed score variance. To provide an estimate of reliability in SEM, therefore, it is necessary to create estimates

**Figure 2**
**Basic Reliability Structural Equation Modeling (SEM) Path Diagram**



of both the true and observed score variances. Using SEM, a measure's total observed score variance can be made available by creating a composite observed variable (*X*). This variable is created by adding the variances of the individual observed variables ($X_1$, $X_2$, etc.) while taking into account the shared variance of the individual observed variables (Miller, 1995; Raykov, 1997a). This process, represented in SEM terms, is shown in Figure 2. Whereas the composite variable *X* is represented as a circle, the direction of the arrows show that *X* is a direct result of the sum of the individual observed item variances, taking into account the variance shared between items.

The creation of an estimate of the true score variance has already been discussed; it is the variance of the latent variable labeled *T* in Figures 1 and 2. To calculate a reliability estimate with this information, one has only to apply Equation 1. Alternatively, an estimate of reliability can be obtained from the model shown in Figure 2, by squaring the implied correlation between the composite latent true variable (*T*) and the composite observed variable (*X*) to arrive at the percentage of the total observed variance that is accounted for by the "true" variable.

## Measurement Models

To identify the measurement model used to estimate the composite true variable, it is necessary to make further assumptions above and beyond unidimensionality. There are a number of measurement models that may be useful in estimating the reliability of test items, each of which requires that the data used meets different requirements.

*The parallel model.* The parallel model is the most restrictive measurement model for use in defining the composite true score. In addition to requiring that all test items measure a single latent variable (unidimensionality), the parallel model assumes that all test items are exactly equivalent to one another. All items must measure the same latent variable, on the same scale, with the same degree of precision, and with the same amount of error (Raykov, 1997a, 1997b). All item true scores are assumed to be equal to one another, and all error scores are likewise equal across items. When applied to the CTT equation, each item *k* for individual *i* can be shown as

$$X_{ik} = T_i + E_i. \tag{2}$$

The parallel model can be used to identify the SEM path diagram shown in Figure 2. To do so, each of the paths from the composite true variable to the individual item variables are set to 1, signifying that each measured variable measures the same latent variable with the same degree of precision and the same scale. Additionally, the individual item error variances are constrained to be equal to one another; in Amos (Arbuckle, 2003), this is accomplished by setting the variance of the error terms to a letter (parameters with the same letter are constrained to equality).

*The tau-equivalent model.* The tau-equivalent model is identical to the more restrictive parallel model, save that individual item error variances are freed to differ from one another. This implies that individual items measure the same latent variable on the same scale with the same degree of precision, but with possibly different amounts of error (Raykov, 1997a, 1997b). All variance unique to a specific item is therefore assumed to be the result of error. The tau-equivalent model implies that although all item true scores are equal, each item has unique error terms:

$$X_{ik} = T_i + E_{ik}. \tag{3}$$

The SEM path diagram for the tau-equivalent model is identical to the parallel model path diagram, save that error variances are no longer constrained to equality.

*The essentially tau-equivalent model.* The essentially tau-equivalent model is, as its name implies, essentially the same as the tau-equivalent model. Essential tau-equivalence assumes that each item measures the same latent variable, on the same scale, but with possibly different degrees of precision (Raykov, 1997a). Again, as with the tau-equivalent model, the essentially tau-equivalent model allows for possibly different error variances.

The difference between item precision and scale is an important distinction to make. Whereas tau-equivalence assumes that item true scores are equal across items, the essentially tau-equivalent model allows each item true score to differ by an additive constant unique to each pair of variables (Miller, 1995; Raykov, 1997a). Mathematically, this assumption can be represented as shown:

$$X_{ik} = (\alpha_{\kappa} + \mathrm{T}_{i}) + E_{ik}. \tag{4}$$

These equations reflect the fact that, although items' true scores are being measured on the same "scale" (similar variances), they may differ in terms of "precision" (different means). A "precise" measure would be one in which the measured values for different items would be closely grouped together, whereas the measured values of different items would be widely spread out in an "imprecise" measure. For example, consider a test designed to measure the latent variable depression in which each item is measured on a 5-point Likert-like scale, from *strongly disagree* to *strongly agree*. Responses to the items "I feel sad sometimes" and "I almost always feel sad" are likely to share similar distributions, though perhaps with different modes. This might be due to the fact that, though both questions measure the same latent variable on the same scale, the second question is worded more strongly than the first. As long as the variances of these questions are similar across respondents, they are both measuring depression in the same scale, whereas their precision in measuring depression differs.

The inclusion of an additive constant affects only an item's mean, not its variance or covariances with other items. As reliability is a variance-accounted-for statistic, it is unaffected by differing means. Therefore, for the purposes of estimating reliability, the SEM path diagram for the essentially tau-equivalent model is identical to that of the tau-equivalent model.

It should be noted that coefficient alpha, the most widely used estimate of internal consistency, is an estimate of reliability based on the essentially tau-equivalent model. Using SEM procedures with the essentially tau-equivalent measurement model will result in a reliability estimate that is equal to Cronbach's alpha. Because it is based on the essentially tau-equivalent model, Coefficient alpha assumes that all items measure the same latent trait on the same scale, with the only variance unique to an item being comprised wholly of error.

*The congeneric model.* The congeneric model is the least restrictive, most general model of use for reliability estimation. The congeneric model assumes that each individual item measures the same latent variable, with possibly different scales, with possibly different degrees of precision, and with possibly different amounts of error (Raykov, 1997a). Whereas the essentially tau-equivalent model allows item true scores to differ by only an additive constant, the congeneric model assumes a linear relationship between item true scores, allowing for both an additive and a multiplicative constant between each pair of item true scores, as shown below:

$$X_{ik} = [\alpha_{k} + \beta_{k}(\mathrm{T}_{i})] + E_{ik}. \tag{5}$$

To identify the model shown in Figure 2 with the congeneric model, the path from the latent true variable to one of the measured items is set to 1, whereas the other paths from the true variable to the items are left free to be estimated. This indicates that the true scores of the other items are expressed in terms of the true score of the fixed item. Any of the measured items can be chosen as the "scaling" variable, with no effect on the outcome of the model.

## The Hierarchical Nature of Measurement Models

Each of the previously discussed measurement models is part of a nested hierarchy. The congeneric model shown in Equation 5 is the most general, least restrictive model for use in reliability estimation. If all multiplicative constants in the congeneric model are set to 1 (inferring that item true scores are measured on the same scale, or have the same standard deviation), we arrive at the essentially tau-equivalent, or coefficient alpha, model shown in Equation 4. If all additive constants are then set to 0 (inferring that not only do item true scores have the same variance, but they are measured with the same degree of precision, or have the same mean), we arrive at the tau-equivalent model shown in Equation 3. Finally, if all error variances are set to equal one another, we arrive at the parallel model shown in Equation 2, where all observed and latent variables are equivalent across items.

## Estimating Reliability With the Hierarchical Model

Coefficient alpha is considered a lower bound estimate of reliability, and the extent of coefficient alpha's underestimation of the reliability cannot be typically known. One common reason for coefficient alpha's underestimation of reliability is the violation of the essentially tau-equivalent model. For example, if data characteristics indicate that the test items measure the same latent variable in different scales, coefficient alpha (based on the essentially tau-equivalent model) would underestimate the reliability, which would be better estimated using the congeneric model. Using the appropriate model given the characteristics of the data can provide a much more accurate estimation of reliability. As the measurement models used in estimating reliability are hierarchical, the fit of the data can be tested to each model in a step-by-step manner, working from least restrictive/parsimonious to most restrictive/parsimonious. This allows the assumptions of the reliability estimates to be tested and the best possible model chosen.

The process of determining which measurement model to use to estimate reliability is a simple one. First, the fit of the congeneric model is tested, and fit statistics are obtained. Next, the fit of the essentially tau-equivalent/tau-equivalent model is tested, and the resulting fit statistics are compared to those obtained from the congeneric model. If the decrease in fit is large enough to be considered meaningful (by whatever means the individual researcher chooses—fit statistics, $\chi^2$ change, etc.), the congeneric model is used to estimate reliability. If the fit statistics are similar, the fit of the tau-equivalent model is compared to the fit of the parallel model. If the difference in fit between the tau-equivalent and parallel models is meaningful, the tau-equivalent model is used. If the differences are not meaningful, the parallel model is used.

The fit of the congeneric model will always be better than or equal to all other models, as the congeneric model is the least restrictive. For the sake of parsimony, however, the most restrictive feasible model should always be used. The method of determining whether difference in fit between two competing models is meaningful is largely a matter of what fit statistics a given researcher prefers. For the following examples, the goodness-of-fit index (GFI; Jöreskog & Sörbom, 1984), the comparative fit index (CFI; Bentler, 1990), and the root mean square error of approximation (RMSEA; Steiger & Lind, 1980) are used in conjunction with change in $\chi^2$ statistics to evaluate model fit.

It should be noted that the $\chi^2$ goodness-of-fit test statistic utilizes traditional statistical significance testing procedures and is therefore highly subject to the size of the sample being used. As stated by Bentler and Bonett (1980),

> In very large samples virtually all models that one might consider would have to be rejected as statistically untenable. . . . This procedure cannot be justified, since the chi-square value . . . can be made small simply by reducing the sample size. (p. 591)

Although the use of the $\chi^2$ statistic alone for the purpose of determining model fit is questionable, $\chi^2$ statistics can be of use in comparing the fits of nested models in which the sample size is held constant across models (Thompson, 2004).

Another important consideration in the examples that follow is the method of estimation used. Coefficient alpha and the majority of commonly used statistical procedures use ordinary, or unweighted, least squares (OLS), a method of estimation that maximizes explained variance while minimizing unexplained, or error, variance. Maximum likelihood, another method of estimation, attempts to maximize the fit of the data to a given model. As each method of estimation serves a different function, both are used in the following examples. Maximum likelihood is used to initially test the fit of the data to the different models, and OLS is used to provide a reliability estimate comparable to (or, in the case of the tau-equivalent model, exactly equal to) coefficient alpha.

*Heuristic example.* As a demonstration of how coefficient alpha underestimates score reliability when the assumption of essential tau-equivalence is violated, consider the following example. This example uses a fictional five-item measure with 60 individuals. To create the items, a true score was first created for each individual by arbitrarily entering numbers from a computer number pad. These true scores ranged from 1 to 9, with a mean of 5.15 and a standard deviation of 2.11. Error terms were then created for items to vary within individuals and to be completely uncorrelated with one another. An additive constant was then created for each item so that the constants varied from item to item but were constant across participants. The additive constants were arbitrarily set to 1 for $x_1$, 4 for $x_2$, 9 for $x_3$, 5 for $x_4$, and 3 for $x_{5.}$ It should be noted that the inclusion of the additive constants does not impact the results of these analyses, as reliability is dependent on score variance and measures of dispersion are not impacted by additive constants. In regards to reliability, essential tau-equivalence and

**Table 1**
**Variance/Covariance Matrix for Heuristic Data**

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ |
|---|---|---|---|---|---|---|---|
| $x_1$ | 4.98 | | | | | | |
| $x_2$ | 4.60 | 5.59 | | | | | |
| $x_3$ | 4.45 | 4.42 | 6.30 | | | | |
| $x_4$ | 3.84 | 3.81 | 3.66 | 6.44 | | | |
| $x_5$ | 5.71 | 5.67 | 5.52 | 4.91 | 11.86 | | |
| $x_6$ | 23.85 | 23.68 | 22.92 | 19.87 | 34.28 | 127.65 | |
| $x_7$ | 46.53 | 46.20 | 44.67 | 38.57 | 62.30 | 244.36 | 471.95 |

tau-equivalence are essentially the same thing. Item scores ($x_1, x_2, \ldots, x_5$) were created by applying Equation 4. This procedure created a data set that perfectly meets the assumption of essential tau-equivalence.

To test the impact of the violation of the assumption of tau equivalence on reliability, two additional variables were created. These variables ($x_6$ and $x_7$) were created using the same error scores and initial true scores as Item $x_5$; however, the true scores for $x_6$ and $x_7$ differed from the true scores for Items $x_1$ through $x_5$ by a multiplicative constant. These items were created using Equation 5, where $b_6 = 5$ and $b_7 = 10$. This resulted in items that were congeneric to the original five items but violated tau-equivalence. The variance/covariance matrix for items $x_1$ through $x_7$ is presented in Table 1.

The following analyses were conducted using Amos (Arbuckle, 2003), an SEM software program with a graphical, user-friendly interface. Initially, Items $x_1$ through $x_5$ were subjected to a reliability analysis using the tau-equivalent measurement model. Although the correlation between $X$ and $T$ was not explicit in the model, Amos allows one to select "standardized estimates" and "all implied moments" as output options. This produces a correlation matrix between all latent and observed variables included in the model. The correlation between $X$ and $T$ using OLS as a method of estimation was squared to provide a reliability estimate. As shown in Table 2, this resulted in a tau-equivalent reliability of .91. When this same set of item scores was subjected to a reliability analysis using a congeneric measurement model, it also resulted in a reliability estimate of .91. Because the items scores are perfectly tau-equivalent to one another, and because the tau-equivalent model is a special case of the congeneric model, the reliability estimates obtained by either method are identical.

Because the tau-equivalent and congeneric measurement models are nested models, the difference in fit of these two models can be obtained by using maximum likelihood as the method of estimation and looking both at the differences in fit indices and at the change in $\chi^2$, which is obtained by simply subtracting the $\chi^2$ and degrees of freedom of the congeneric model from the tau-equivalent model. The fit indices shown in Table 2 show that both the tau-equivalent and congeneric models provide an excellent fit to the data, with CFIs and GFIs greater than .9 and RMSEA values less than .08. In

**Table 2**
**Model Fit and Reliability Estimates for Heuristic Data**

| Item | Model | Reliability | GFI | CFI | RMSEA | $\chi^2$ | *df* | *p* |
|---|---|---|---|---|---|---|---|---|
| $x_1$-$x_5$ | Tau-equivalent | .91 | .966 | 1.000 | .000 | 5.5 | 9 | |
| | Congeneric | .91 | .999 | 1.000 | .000 | 0.1 | 5 | |
| | Change | | | | | 5.4 | 4 | .249 |
| $x_1$-$x_4$, $x_6$ | Tau-equivalent | .76 | .728 | .630 | .473 | 127.8 | 9 | |
| | Congeneric | .97 | 1.000 | 1.000 | .000 | < .1 | 5 | |
| | Change | | | | | 127.8 | 4 | < .001 |
| $x_1$-$x_4$, $x_7$ | Tau-equivalent | .56 | .715 | .556 | .537 | 161.9 | 9 | |
| | Congeneric | .99 | 1.000 | 1.000 | .000 | < .1 | 5 | |
| | Change | | | | | 161.9 | 4 | < .001 |

Note: GFI = goodness-of-fit index; CFI = comparative fit index; RMSEA = root mean square error of approximation.

fact, given that the data were constructed to fit the model, the fit indices indicate near-perfect fit. Additionally, the difference in $\chi^2$ between these two models is neither large nor statistically significant. Both models fit the data equally well; therefore, one might choose to select the more restrictive, parsimonious tau-equivalent estimate over the congeneric estimate.

Next, Item $x_5$ was replaced by Item $x_6$ and the same sets of analyses were run again. Item $x_6$ is identical to Item $x_5$, save that the true score of $x_6$ differs from the true score of $x_5$ by a multiplicative constant of 5. As seen in Table 2, the tau-equivalent measure of reliability (.76) is substantially lower than the congeneric measure of reliability (.97). The fit indices suggest that the data fits the congeneric model rather well but has a poor fit with the tau-equivalent model. Additionally, the $\chi^2$ difference in fit between these models is both large and statistically significant; therefore, the congeneric model appears to be the best fit for the data. Had one simply used Cronbach's alpha without testing whether the data met the tau-equivalence assumption, one would have underestimated the reliability of the item scores.

Finally, Table 2 also shows the results of the reliability analyses using $x_7$ instead of $x_5$. The true score of $x_7$ differs by a multiplicative constant of 10 from the true score of Item $x_5$. Again, the fit indices and the $\chi^2$ difference test agree that the data best fit the congeneric model. As shown, were one to use Cronbach's alpha, one might erroneously assume that a near-perfectly reliable congeneric measure is made up of almost half error variance.

## Factors Affecting Coefficient Alpha's Underestimation of Reliability

As demonstrated, coefficient alpha underestimates the reliability of test scores when the test violates the assumption of tau-equivalence. Specifically, the larger the violation of tau-equivalence that occurs, the more coefficient alpha underestimates score reliability. Both the present example and previous work (Raykov, 1997b) have demonstrated that the presence of even a single item that is not tau-equivalent to the other items can have a dramatic impact on the accuracy of coefficient alpha; however, the impact that violating the assumption of tau-equivalence can have is also dependent on a number of other factors.

All other things being considered equal, tests with a greater number of items are less vulnerable to underestimation when tau-equivalence is violated than tests with only a small number of items (Raykov, 1997b). This is due to the fact that, when a single item violates tau-equivalence, the proportion of true score variance that is congeneric to the other item true scores is smaller when one has a greater number of items than when one has fewer items.

Because the tau-equivalence model assumes that items are measured on the same scale, examining item standard deviations may be of some utility. If the standard deviations of item scores composing a test are vastly different from one another, they are likely to be being measured on different scales. Such a comparison might be made by constructing confidence intervals about item standard deviations, and visually examining them for equivalence. If the items standard deviations were not equivalent, one might be alerted that the data may be failing the assumption of tau-equivalence. It should be noted that the standard deviation of an item could be impacted by both variance in the true score and variance in the error term associated with the item. An examination of the standard deviations of item true scores (as opposed to item observed scores) would give a better estimate of the degree to which tau-equivalence is violated. Because item true scores are not typically known, calculating the standard deviations of the item true scores is not likely to be feasible.

Finally, tests that use multiple response formats across items are more likely to violate the assumption of tau-equivalence than those that do not. For example, the item true scores from several true-false items are likely to be measured on a different scale than the true scores of items that are scored on a 6-point Likert-type scale. The use of different response formats typically indicates that the items are being measured on different scales, and is likely to result in different true-score standard deviations from item to item.

*Applied example.* Whereas the heuristic example demonstrated how coefficient alpha underestimates reliability when the assumptions of tau-equivalence are violated, an example using actual data may also be instructive in providing an example of how a confluence of the previously discussed factors can impact the accuracy of coefficient alpha. The following data are from a study conducted by Graham and Conoley (2006) using the Dyadic Adjustment Scale (DAS; Spanier, 1976). The DAS is a commonly used measure of relationship quality, designed for use with cohabiting couples. The DAS is a 32-item self-report measure in a variety of response formats whose sum results in a number from 0 to 151, with a higher number denoting greater relationship quality. Whereas the total score from the DAS is most often used in applied research, the test developer originally divided the scale into four subscales. One of these subscales, Affective Expression, consists of 4 items inquiring about levels of agreement between

**Table 3**
**Variance/Covariance Matrix for Dyadic Adjustment Scale (DAS) Husband Data From Graham and Conoley (2006)**

| | $DAS_4$ | $DAS_6$ | $DAS_{29}$ | $DAS_{30}$ |
|---|---|---|---|---|
| $DAS_4$ | .903 | | | |
| $DAS_6$ | .657 | .986 | | |
| $DAS_{29}$ | .042 | .182 | .236 | |
| $DAS_{30}$ | .199 | .209 | .072 | .173 |

spouses on demonstrations of affection and sex relations (both on 6-point Likert-type scales) and about whether the couple has disagreed recently as a result of being too tired for sex and not showing love (both dichotomously scored yes–no). A recent reliability generalization meta-analysis of the DAS indicated a mean Affective Expression subscale score reliability of .71 across studies (Graham, Liu, & Jeziorski, in press). Table 3 presents the variance/covariance matrix of the 4 affective expression items of 60 husbands from the study conducted by Graham and Conoley.

Initially, one might notice that the number of items comprising this subscale is relatively small; this makes the alpha reliability of scores from these items more vulnerable to violations of essential tau-equivalence. One might next notice that two of the items are scored dichotomously, whereas the other two items are scored on a 6-point Likert-type scale. This use of different response formats makes violation of the assumption of tau-equivalence more likely. Next, one might notice that the variances of the two dichotomously-measured items appear to be substantially lower than the two items scored on a 6-point scale. Confidence intervals constructed about the standard deviations of Items 4 (.95 ± .15) and 6 (.99 ± .15) do not overlap with the confidence intervals constructed about the standard deviations of Items 29 (.47 ± .07) and 30 (.42 ± .06).

In all instances, this measure gives the appearance of violating the assumption of essential tau-equivalence. As such, it would be expected that coefficient alpha would provide a lower estimate of reliability than a congeneric measure. Using both SPSS and the tau-equivalent SEM model described above, these data have a Cronbach's alpha of .72. Following the procedures for estimating congeneric reliability, an estimate of .83 is obtained. The data fit the congeneric model (GFI = .925; CFI = .865; RMSEA = .288) better than the tau-equivalent model (GFI = .761; CFI = .463; RMSEA = .363). Additionally, the congeneric model had a statistically significantly better fit than the tau-equivalent model, $\Delta\chi^2(3) = 32.1$, $p < .001$. Across studies, the Affective Expression subscale of the DAS has been shown to consistently result in scores with lower reliability than scores on the other subscales (Graham et al., in press). It appears that this underestimation may be in part due to the fact that these scores violate the assumption of tau-equivalence. In this example, Cronbach's alpha underestimated the reliability of these scores by at least 10%, because the subscale better fits a congeneric model!

## Discussion

The use of coefficient alpha to estimate reliability is often taken for granted, with little thought put into understanding the assumptions required for the obtained reliability estimate to be accurate. As a result, the reliability of published data is often needlessly underestimated. Nearly all classical general linear procedures require that certain assumptions be met (normality, homogeneity of variance, etc.), yet these procedures are routinely applied even when the basic assumptions have not been met (Wilcox, 1998). Wilkinson and the American Psychological Association (APA) Task Force on Statistical Inference (1999) addressed this by stating simply that "you should take efforts to assure that the underlying assumptions for the analysis are reasonable given the data" (p. 598). Although this statement may seem overly simplistic, the fact remains that many students and researchers in education and psychology are unaware of many of the assumptions required by a given statistical procedure, are unaware of how to test those assumptions, or are unaware of acceptable alternatives should those assumptions not be met.

A survey of graduate programs in psychology reported that only 27% of programs reported that most or all of their students can appropriately apply methods of reliability measurement to their own research (Aiken et al., 1990). Aiken and colleagues (1990) concluded that this deficiency in understanding basic measurement concepts "opens the door to a proliferation of poorly constructed ad hoc measures, potentially impeding future progress in all areas of the field" (p. 730). The assumption of essentially tau-equivalence in reliability is rarely discussed out of measurement and SEM circles, and even the measurement literature rarely (if ever) considers the assumptions of essentially tau-equivalence when reporting reliability. The case of tau-equivalence in the field of measurement is particularly perplexing when one considers other procedures commonly used in the development of measures. Whereas the exploratory and confirmatory factor analytic procedures used in determining measure item composition assume the more general case of congeneric items, the statistic most often used to estimate the reliability of the resultant item groupings assumes tau-equivalence. In short, the factor models and measurement models do not match!

The procedures discussed here can be easily replicated using Amos or any other commonly available SEM software package. The use of SEM techniques does, however, require large sample sizes. Although this is likely to limit this technique's everyday use, it is amenable to the vast majority of psychometric studies which typically use large sample sizes. Certainly, examining whether one's data meets the assumptions of tau-equivalence should be an important step in the initial development of any measure, as should the calculation of a congeneric estimate of reliability should the data fail that assumption.

The present discussion is not intended to advocate for the use of a congeneric measure of reliability over coefficient alpha in all cases. Measurement, like other areas of science, rewards parsimony. Because the tau-equivalent model estimates fewer parameters than the congeneric model, there are more opportunities to falsify the tau-equivalent model. The more falsifiable a model is, the more rigorous and persuasive it is considered if it is not subsequently falsified (Mulaik et al., 1989). Therefore, if both models fit reasonably well, a tau-equivalent estimate of reliability is preferable over a congeneric estimate. If the congeneric model fits appreciably better, a congeneric estimate of reliability is preferable.

It is hoped that the present discussion provides an accessible framework for understanding and testing the tau-equivalence assumption of Cronbach's alpha and for calculating a congeneric estimate of reliability should that assumption fail. Testing the fit of the various measurement models in a hierarchical fashion is a simple matter given the current availability and ease of use of SEM software packages. Researchers are encouraged to always test whether their data are essentially tau-equivalent when developing a measure and to use the appropriate measurement model to estimate reliability. Researchers with sufficiently large samples are also encouraged to utilize these procedures, even if the psychometric properties of their data are not the primary focus. This is particularly important if a small number of items compose the measure, if items utilize different formats, or if items have largely different standard deviations. The use of these procedures can help ensure that the assumptions required by commonly used reliability estimates are not violated and that an accurate estimate of reliability is obtained.

## References

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (with Roediger, H. L., Scarr, S., Kazdin, A. E., & Sherman, S. J.). (1990). The training in statistics, methodology, and measurement in psychology. *American Psychologist, 45,* 721-734.

Arbuckle, J. L. (2003). Amos (Version 5.0.1) [Computer software]. Spring House, PA: Amos Development Corporation.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107,* 238-246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). Phoenix, AZ: Ornyx.

Graham, J. M., & Conoley, C. W. (2006). The role of marital attributions in the relationship between life stressors and marital quality. *Personal Relationships, 13,* 231-244.

Graham, J. M., Liu, Y. J., & Jeziorski, J. L. (in press). The Dyadic Adjustment Scale: A reliability generalizability meta-analysis. *Journal of Marriage and Family.*

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI user's guide* (3rd ed.). Mooresville, IN: Scientific Software.

Lord, F. M., & 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 perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling, 2,* 255-273.

Mulaik, S. A., James, L. R., van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 195,* 430-445.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum.

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

Raykov, T. (1997b). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32,* 329-353.

Spanier, G. B. (1976). Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads. *Journal of Marriage and Family, 38,* 15-28.

Steiger, J. H., & Lind, J. C. (1980, June). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychonomic Society, Iowa City, IA.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics *is* datametrics: The test is not reliable. *Educational and Psychological Measurement, 60,* 174-195.

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60,* 509-522.

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement, 58,* 21-37.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300-314.

Wilkinson, L., & American Psychological Association (APA) Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.