ITEMS Module Companion Article: Scale Reliability in Structural Equation Models

Content Developers: Gregory R. Hancock & Ji An, *University of Maryland*
ITEMS Portal Editor: André A. Rupp, *Educational Testing Service*

In this ITEMS module we frame the topic of scale reliability within a confirmatory factor analysis and *structural equation modeling* (SEM) context to address some of the limitations of Cronbach's α. This modeling approach has two major advantages: (1) it allows researchers to make explicit the relation between their items and the latent variables representing the constructs those items intend to measure, and (2) it facilitates a more principled and formal practice of scale reliability evaluation. Specifically, we begin the module by discussing key conceptual and statistical foundations of the *classical test theory* model and then framing it within an SEM context; we do so first with a single item and then expand this approach to a multi-item scale. This allows us to set the stage for presenting different measurement structures that might underlie a scale and, more importantly, for assessing and comparing those structures formally within the SEM context. We then make explicit the connection between measurement model parameters and different measures of reliability, emphasizing the challenges and benefits of key measures while ultimately endorsing the more flexible McDonald's ω over Cronbach's α. We then demonstrate how to estimate key measures in both a commercial software program (M*plus*) and three packages within an open-source environment (*R*). In closing, we make recommendations for practitioners about best practices in reliability estimation based on the ideas presented in the module.

Keywords: scale reliability; classical test theory; structural equation modeling; Cronbach's α; McDonald's ω; parallel model; tau-equivalent model; congeneric model; relative model-data fit; M*plus*; *R*.

For content-related issues please contact Gregory R. Hancock at 1230 Benjamin Building, 3942 Campus Drive, University of Maryland, College Park, MD 20742-1115; Phone: 301-405-3621; e-mail: ghancock@umd.edu. For editorial issues please contact André A. Rupp at 660 Rosedale Road, Mailstop T-03, Princeton, NJ 08541; Phone: 609-252-8545; e-mail: arupp@ets.org.

Prerequisite Knowledge

This ITEMS module assumes that learners have had some exposure to basic principles of classical test theory (CTT) and reliability even though we provide a brief treatment of such concepts specifically within the SEM framework. Specifically, we assume that learners are familiar with:

- true scores and error scores;
- definitions of reliability in CTT;
- Cronbach's α.

Specific prior computational experience with specifying structural models in M*plus* or *R*, while helpful in working through examples, is not critical as we provide instructional scaffolds throughout.

Learning Objectives

Upon completion of this ITEMS module, learners should be able to:

A. Conceptual Understanding

- Express a unidimensional scale graphically as a structural model
- Express parallel, tau-equivalent, and congeneric models within an SEM framework
- Express the assumptions of Cronbach's α in terms of a structural model
- Determine, for any given unidimensional scale, which parameters would exist for the parallel model, tau-equivalent model, and congeneric model

B. Working with Software

- Fit a unidimensional congeneric model
- Compute the necessary summary statistics to obtain the traditional estimate of Cronbach's α
- Compute an estimate of Cronbach's α using parameter estimates from output
- Compute an estimate of McDonald's ω using parameter estimates from output
- Compute a 95% confidence interval for McDonald's ω using the asymptotic standard error (software permitting)
- Compute a 95% bootstrap confidence interval for McDonald's ω (software permitting)
- Conduct absolute model-data fit evaluations using suitable indices (e.g., χ^2, AIC, BIC, SRMR, RMSEA, CFI) for parallel, tau-equivalent, and congeneric models
- Conduct relative model-data fit comparisons among the parallel, tau-equivalent, and congeneric models using information criteria as well as χ^2 difference tests

After completion of this module, learners might take additional ITEMS modules on CTT, generalizability theory, validity, item response theory, and bi-factor models. Check the NCME webpage for up-to-date information on ITEMS modules.

All too often the areas of measurement, statistics, assessment, and evaluation are treated as separate entities, collectively contributing to a well-rounded quantitative education but also serving to define specializations within the broader quantitative domain. Although such divisions, and their own subdivisions, have existed and continue to persist largely for historical reasons, they have also become increasingly artificial, unnecessary, and even to some extent detrimental.

Generally speaking, all of these areas utilize models that serve as explanatory place-holders for processes, and sometimes for variables within those processes, that are otherwise difficult, if not impossible, to observe directly. Whether these models represent the unobserved (i.e., latent) variables hypothesized to underlie response patterns to specific psychological or achievement instruments, individual differences in change in key outcomes over time, or the complex interplay of the characteristics of test items, test takers, and testing contexts, all nonetheless distill down to the same core elements: *variables* and *links*.

The former may include *measured variables* for which we have direct observations/data as well as *latent variables* for which we have strong theory but no direct observations, while the latter constitute the hypothesized connections within and between both types of variables, such as linear or logistic, and occasionally links between the variables and other links, as in cases of moderation. In short, models are models, no matter what their historical origin or typical application.

The existence of the aforementioned historical divisions has meant that different areas have advanced more than others in some respects, less than others in other respects, with their respective evolutions guided by the goals and needs of each specific area. On the positive side, this reflects the fact that areas are growing and adapting to meet their specific needs. On the negative side, however, it also means that wheels are often reinvented under different names, when ways of thinking and doing that are much needed in one area might have existed right "next door" for quite some time.

The topic of scale reliability is one born out of *classical test theory* (CTT), and which has evolved primarily within the measurement/psychometric domain (see, e.g., Allen & Yen, 1979; Crocker & Algina, 1986). As this ITEMS module will show, it is also a topic that can benefit from being framed within a latent variable context, one most typically associated with *confirmatory factor analysis* (CFA) and *structural equation modeling* (SEM) (see, e.g., Kline, 2016). The benefits of such a framing will be primarily two-fold, the first pedagogical and the second methodological.

Framing scale reliability within an SEM context will first serve to clarify the roles of measured scale items and the latent constructs they intend to measure, make explicit the links between the items and their constructs, and, in turn, formalize what we mean by *scale reliability* and the typical indices thereof. Second, it will allow us to conduct scale reliability analysis in a more principled and formal way, assessing and possibly remediating the models that underlie reliability measures and accommodating real-world data challenges along the way.

This ITEMS module is structured as follows. Specifically, we begin the module by discussing key conceptual and statistical foundations of the CTT model and then framing it within an SEM context; we do so first with a single item and then expand this approach to a multi-item scale. This allows us to set the stage for presenting different measurement structures that might underlie a scale and, more importantly, for assessing and comparing those structures formally within the SEM context. We then make explicit the connection between measurement model parameters and different measures of reliability, emphasizing the challenges and benefits of key measures while ultimately endorsing the flexible McDonald's ω over Cronbach's α. We then demonstrate how to estimate key measures in both a commercial software program (M*plus*) and three packages within an open-source environment (*R*). In closing, we make recommendations for practitioners about best practices in reliability estimation based on the ideas presented in the module.

Conceptual Foundations

Whether for evaluating attitudes, beliefs, or achievement, scales are typically composed of *items* whose purpose is to provide insight into the underlying *construct(s)* they are designed to measure. For the goals of the current ITEMS module, we will primarily consider scales or subscales designed to be *unidimensional*, that is, whose items are intended to reflect a single latent construct.

As an example, Midgley et al. (1998) created an instrument with 18 scale items that are intended to tap three constructs for school-age children for assessing achievement goal orientations, with subsets of 6 items dedicated to each of the constructs "Ability-Approach Goal Orientation," "Ability-Avoid Goal Orientation," and "Task Goal Orientation." For the purposes of this ITEMS module, imagine that we have data on the "Task Goal Orientation" subscale for which the items are:

X_1: I like school work that I'll learn from, even if I make a lot of mistakes.

X_2: An important reason why I do my school work is because I like to learn new things.

X_3: I like school work best when it really makes me think.

X_4: An important reason why I do my work in school is because I want to get better at it.

X_5: I do my school work because I'm interested in it.

X_6: An important reason I do my school work is because I enjoy it.

Subjects rate themselves on a 7-point rating scale for each item, which ranges from 1 = 'not true of me at all' to 7 = 'very true of me.' The practitioner's ultimate goal for these six items is typically to get a total scale score, which can in turn be used as a measured proxy for students' latent "Task Goal Orientation" with a degree of reliability that can be assessed.

Before addressing the reliability of the scale as a whole, however, let us start by considering each scale item individually. As per CTT, we may view each *i*th individual's observed X_i score on a scale item as being composed of that individual's true score T_i and error E_i, where T_i is the long-run expected value of X_i for that individual upon theoretically infinite re-administrations of item X, and E_i is, quite simply, the remainder (see, e.g., Traub & Rowley, 1991):

$$X_i = T_i + E_i\ . \qquad (1)$$

If one is willing to believe that T is not only the long-run expected value of X, but, in fact, represents a theoretical score on an underlying construct of interest, such as true "Task Goal Orientation," and that E is a constellation of other agents that also contribute to one's observed score on a given variable/item X, then Equation 1 may be represented graphically as in Figure 1.

In Figure 1 X is depicted inside a box, indicating that it is a measured variable, while T is depicted inside a circle indicating that it is not directly observed (i.e., latent); we note that E could also be contained within a circle due to its unobserved nature, but doing so is less customary. Additionally, arrows are used to reflect a belief that both T and E contribute to, and indeed influence, scores observed on a given X; this link is typically approximated as a linear relation and we will assume so here.

Furthermore, the paths from T and E into X are set to 1, indicating that X receives input from each in the same metric as X. This does not imply that they inform X equally, however. Indeed, T and E each have their own variance contributing to that of X, as indicated by the attached and labeled two-headed arrows; these variances in turn serve the traditional definition of reliability:

$$\text{rel}(X) = \frac{\text{var}(T)}{\text{var}(X)} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}\ . \qquad (2)$$

Finally, note that T and E are not directly connected. While in CTT this separation reflects the necessity that the error, as a remainder, must be uncorrelated with T (i.e., $E_i = X_i - T_i$), within the modeling world this reflects an explicit or implicit – and occasionally questionable – assumption that the residual influences on X are unrelated to the construct whose score is represented by T.

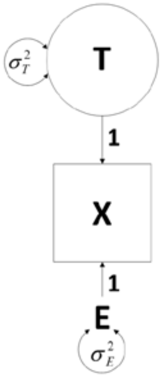


Figure 1 Latent variable representation of the core CTT equation.

After having focused above on a single observed scale item, now let us consider that each such item has its own true score and error score components, and that the observed scores across all scale items correlate because their true scores correlate. Further, and critically, the true scores themselves correlate because they are all believed to be influenced by a single common construct, ξ, such as "Task Goal Orientation."

This latent factor ξ is not the only influence on the true scores, however; each T has influences specific to that particular true score and unrelated to ξ, which we may designate as S (for *specific* factor). T_1, for example, the true score for item X_1, would be influenced by the factor common to all item true scores, ξ , and possibly also by a factor S_1 representing students' attitude about making mistakes, which is independent of their "Task Goal Orientation." Thus, the true score variance shown in Figure 1 is explained by the common factor ξ and the specific factor S_1. Such a model across all scale items is depicted in Figure 2.

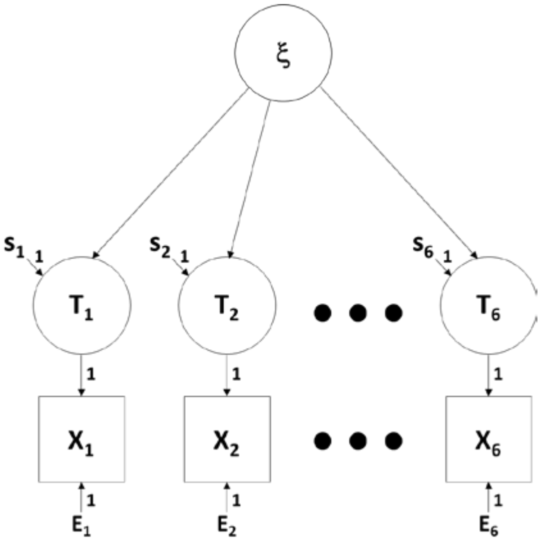


Figure 2 Full latent variable representation of a six-item scale.

It is worth stating that the model in Figure 2 is entirely reasonable from a theoretical point of view; analytically, however, it is generally intractable, suffering from *identification* issues without further assumptions and their associated constraints. Fortunately, we may make a highly useful simplification by decomposing each item's true score into its constituent parts ξ and S, meaning that each measured X variable is, at its core, influenced by the common factor ξ and two independent sources of error, S and E.

assumptions and their associated constraints. Fortunately, we may make a highly useful simplification by decomposing each item's true score into its constituent parts ξ

If ξ were modeled as a direct influence on the X variables, and the error sources were collapsed into a single combined error term δ (representing that which is not explained by the common factor), the resulting model would be a fairly familiar one-factor model depicted in Figure 3.

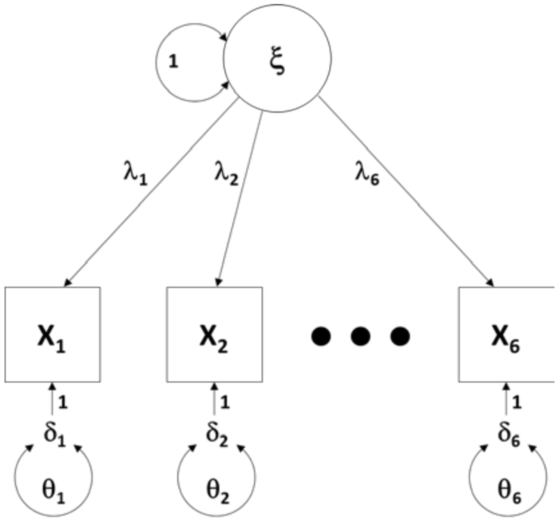


Figure 3 Common factor model representation of a six-item scale.

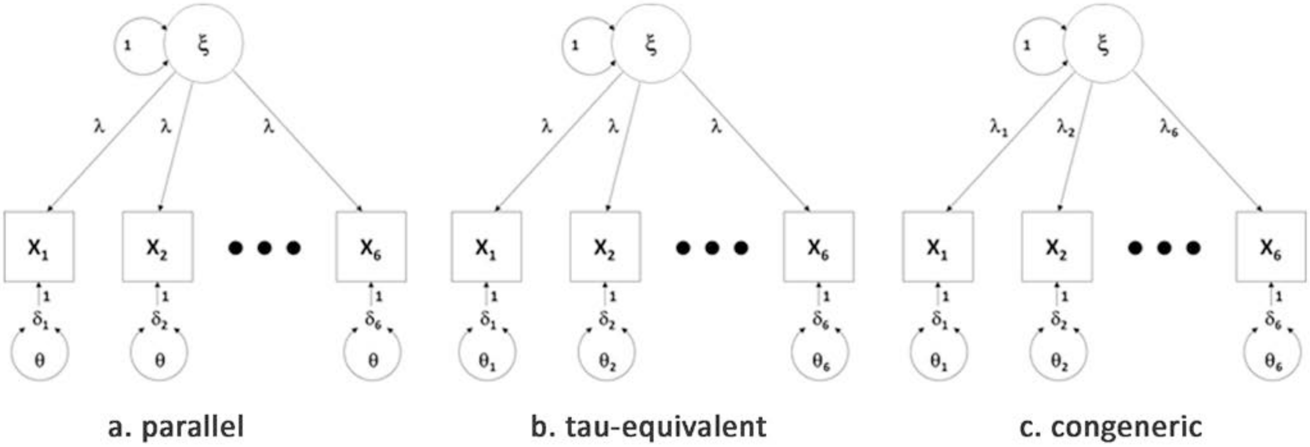
Relative to Figure 2, the true score variance in each original item in Figure 3 has been bifurcated into that which is relevant to the scale as a whole (and thus explainable by the common factor ξ) and that which is specific to the item's true score (i.e., S); this latter component, along with that which is unique to the item's measured score X (i.e., E), are both sources of error (δ) from the perspective of the scale as a whole. The loading paths, reflecting the contribution of the common factor to each scale item, are designated as λ, the variance of each δ error term is θ (a combination of S and E variances) and variance of the factor is set to 1 to identify the factor's metric. The model-based reliability implied for a given scale item X is thus

$$\text{rel}(X) = \frac{\text{variance explained by } \xi}{\text{total item variance}} = \frac{\lambda^2}{\lambda^2 + \theta} \quad , \tag{3}$$

where Equation 3 refers to an item's reliability vis-à-vis the construct being measured by the entire scale, ξ, not the reliability that relates to the item's true score.

Relating the model in Figure 3 back to CTT, if all X variables represent *tests* intended to measure the same construct, a model of *parallel tests* would be one in which all tests have the same amount of true score variance and error variance, σ_T^2 and σ_E^2 in Figure 1, respectively. In order for this to happen, the common factor ξ and specific factor S (see Figure 2) would need to combine to yield the same true score variance for all items, while the item error variances E would likewise need to be the same. Transitioning to Figure 3, then, the implication for the "Task Goal Orientation" scale would be that all six items are explained precisely to the same degree by the common factor $\lambda_1^2 = \ldots = \lambda_6^2 = \lambda^2$ as well as by their aggregate error components δ resulting in equal error variances $\theta_1 = \ldots = \theta_6 = \theta$.

The assumption of parallel tests might be somewhat reasonable for a series of instruments intended to measure, say, a relatively narrowly defined skill such as "single-digit addition" where each test contains random sets of single-digit addition items. The assumption of parallel scale *items*, however, is generally much less palatable, given that specific item content is typically varied within an instrument, as seen in the case of the set of "Task Goal Orientation" items. That is, even if a common factor does underlie all six scale items, it may be hard to believe that the amount of variance explained in X_3 ("I like school work best when it really makes me think") is precisely the same as that explained in X_6 ("An important reason I do my school work is because I enjoy it"). This parallel model, with two parameters to explain the six items' 21 unique variances and covariances (i.e., 19 *df*), is depicted in Figure 4a.



a. parallel

b. tau-equivalent

c. congeneric

Figure 4 Parallel, tau-equivalent, and congeneric models for a six-item scale.

A less restrictive configuration for a set of tests would be that all tests have the same amount of true score variance but that error variances are free to vary across tests; this is referred to as true score equivalence or *tau-equivalent tests*. In the case of the "Task Goal Orientation" scale items, tau-equivalence would represent when the underlying factor explains the same amount of variance in each item, $\lambda_1^2 = \ldots = \lambda_6^2 = \lambda^2$ but with the error variances allowed to be unique across items ($\theta_1, \ldots, \theta_6$).

That is, whereas in the parallel model error variances are assumed the same, thereby implying – because loadings are assumed to be the same as well – that all items have the same observed variance, the tau-equivalent model releases the latter assumption acknowledging that specific *S* and/or *E* factors might realistically account for different amounts of variance in different items on an instrument (see, e.g., Reuterberg & Gustafsson, 1992). Indeed, one would expect this be the case for the "Task Goal Orientation" scale given the diversity of items such as "I like school work best when it really makes me think" and "An important reason why I do my work in school is because I want to get better at it."

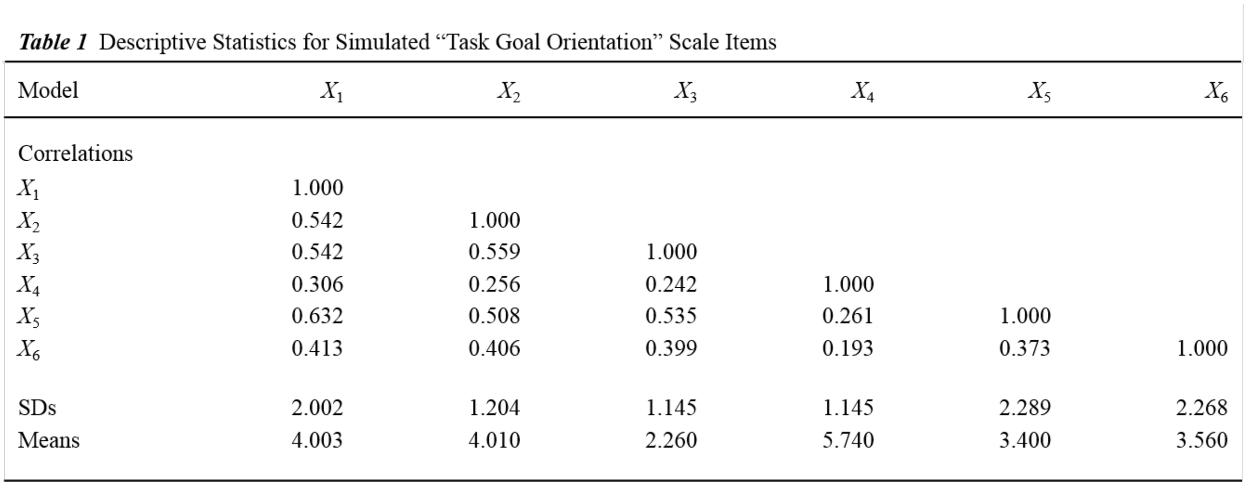
The tau-equivalent scenario, however, while a clear improvement over the highly restrictive parallel scenario, is only slightly more palatable at best, still retaining the generally unrealistic idea that the common factor ξ explains the same amount of variance in all scale items (λ^2). This tau-equivalent model, with seven parameters to explain the six items' 21 unique variances and covariances (i.e., 14 *df*), is depicted in Figure 4b.

Finally, the least restrictive model is one in which the amount of true score variance is allowed to differ across tests as is the amount of error variance; this is referred to as *congeneric tests*. In terms of the set of "Task Goal Orientation" scale items, this translates to potentially different loading paths $\lambda_1, \ldots, \lambda_6$, and potentially different error variances $\theta_1, \ldots, \theta_6$, and, as such, seems the most realistic scenario for a set of scale items all sharing a common underlying factor but still having unique content. This congeneric model, with 12 parameters to explain the six items' 21 unique variances and covariances (i.e., 9 *df*), is represented in Figure 4c.

Having described how to use the SEM framework to characterize the parallel, tau-equivalent, and congeneric models, an immediate advantage is that one need not assume or argue for one set of model conditions or another; instead, we can evaluate model-data fit using standard SEM software packages such as M*plus*, *EQS*, *LISREL*, *AMOS*, or *lavaan*. Results for these models can lead to an evaluation of each model individually via common *absolute fit indices* (e.g., SRMR), *parsimonious fit indices* (e.g., RMSEA), and *incremental fit indices* (e.g., CFI). Models can also be compared relative to one another using *information indices* (e.g., AIC, BIC) and/or using *likelihood ratio tests* with the choice depending on the hierarchical relation among the models that are to be compared. Specifically, likelihood ratio tests are indicated for nested models when a more restrictive model can be expressed as a special case of a more general model whereas information indices are suitable for nested and non-nested models. Based on the previous exposition, the parallel model is nested within the tau-equivalent model and the parallel and tau-equivalent models are nested within the congeneric model.

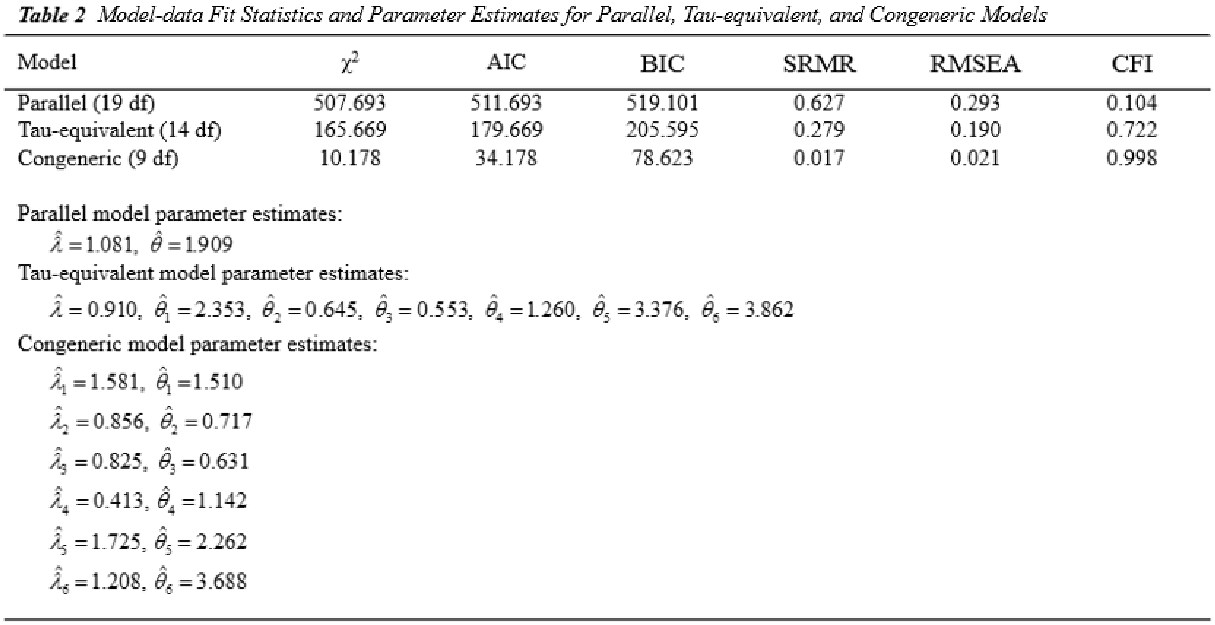
Now let us consider a simulated data set for the six "Task Goal Orientation" items in which 300 subjects were simulated to respond on the aforementioned 7-point rating scale using a congeneric model structure. Descriptive statistics for these data appear in Table 1; the raw data file is available in the online supplementary materials or upon request from the first author.Next, the parallel, tau-equivalent, and congeneric models were fit to the simulated data using M*plus* using the default *maximum likelihood estimation* setting. Syntax for estimating each model appears in Appendix A while model-data fit information and parameter estimates are shown in Table 2.

Table 1 Descriptive Statistics for Simulated "Task Goal Orientation" Scale Items



Model	X_1	X_2	X_3	X_4	X_5	X_6
Correlations						
X_1	1.000					
X_2	0.542	1.000				
X_3	0.542	0.559	1.000			
X_4	0.306	0.256	0.242	1.000		
X_5	0.632	0.508	0.535	0.261	1.000	
X_6	0.413	0.406	0.399	0.193	0.373	1.000
SDs	2.002	1.204	1.145	1.145	2.289	2.268
Means	4.003	4.010	2.260	5.740	3.400	3.560

Table 2 Model-data Fit Statistics and Parameter Estimates for Parallel, Tau-equivalent, and Congeneric Models



Model	χ^2	AIC	BIC	SRMR	RMSEA	CFI
Parallel (19 df)	507.693	511.693	519.101	0.627	0.293	0.104
Tau-equivalent (14 df)	165.669	179.669	205.595	0.279	0.190	0.722
Congeneric (9 df)	10.178	34.178	78.623	0.017	0.021	0.998

Parallel model parameter estimates:

$\hat{\lambda} = 1.081,\ \hat{\theta} = 1.909$

Tau-equivalent model parameter estimates:

$\hat{\lambda} = 0.910,\ \hat{\theta}_1 = 2.353,\ \hat{\theta}_2 = 0.645,\ \hat{\theta}_3 = 0.553,\ \hat{\theta}_4 = 1.260,\ \hat{\theta}_5 = 3.376,\ \hat{\theta}_6 = 3.862$

Congeneric model parameter estimates:

$\hat{\lambda}_1 = 1.581,\ \hat{\theta}_1 = 1.510$

$\hat{\lambda}_2 = 0.856,\ \hat{\theta}_2 = 0.717$

$\hat{\lambda}_3 = 0.825,\ \hat{\theta}_3 = 0.631$

$\hat{\lambda}_4 = 0.413,\ \hat{\theta}_4 = 1.142$

$\hat{\lambda}_5 = 1.725,\ \hat{\theta}_5 = 2.262$

$\hat{\lambda}_6 = 1.208,\ \hat{\theta}_6 = 3.688$

As one would expect given the simulation setup, only the congeneric model meets currently accepted standards of model-data fit for absolute, parsimonious, and incremental indices (for guidelines see, e.g., Mueller & Hancock, 2010). This model would also be the model of choice using both information and statistical criteria: AIC and BIC values are lowest for the congeneric model. Finally, using the likelihood ratio test one can see that the congeneric model fits statistically significantly better than either of the other two models as well (congeneric vs. parallel χ^2_{diff} = 497.515, df=10, p<.001; congeneric vs. tau-equivalent χ^2_{diff} = 155.491, df=5, p<.001). As a result, parameter estimates from the congeneric model for further analyses and reporting.

Reliability Estimation

Having laid the foundations for modeling the formal structure governing the items of a unidimensional instrument, we can now depict the scale itself within the model. To start, given that equally-weighted item aggregation (e.g., computing simple sum scores or averages) is far and away the most common method for deriving a scale score intended to approximate the underlying factor, we will graphically represent an instrument that is the simple sum of its scale items.

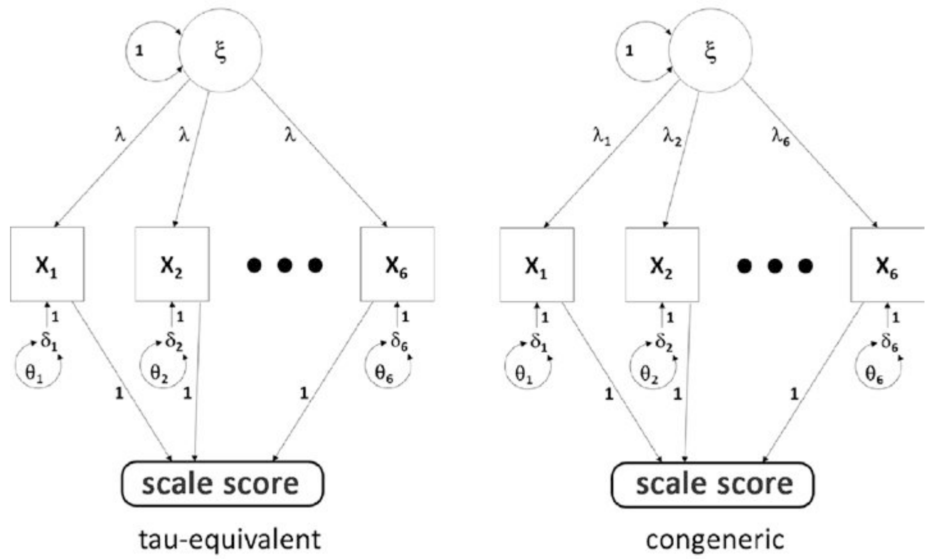


Figure 5 Tau-equivalent and congeneric models with scale scores.

We could do so in any of the models depicted in Figure 4; for our purposes, we will augment the tau-equivalent and congeneric models, as seen in Figure 5, with an entity representing the simple (unit-weighted) sum scale score. This score is depicted as a round-edged rectangle; this is not standard representation per se, but is merely meant to reflect that this is neither measured (rectangle) nor latent (circle) as far as the model is concerned. Indeed, if we had computed this scale score and included it as a measured variable within this model, its perfect multicollinearity with the individual items would have prevented the model from being estimable. It is therefore not literally included, but only graphically depicted as such, and without any residual variance of its own given that it is perfectly determined by its own scale items.

We may now consider reliability in the context of this model. Specifically, we consider the reliability of the scale score that is perfectly determined by the scale items, which themselves are partially determined by their common underlying factor. Said differently, the variance in the scale score has two types of sources, one associated with the factor that it will be used to approximate and one associated with the item-level error terms.

For the tau-equivalent model, using either the linear algebra of composites or heuristically equivalent path tracing (see, e.g., Loehlin, 2004), the total variance in the scale score for a *J*-item instrument may be shown to be:

$$\begin{aligned}&\text{var(tau-equivalent scale)}\\&= \text{variance explained by } \xi\\&+ \text{variance explained by error terms}\\&= J^2\lambda^2 + \sum_{i=1}^{J}\theta_i \ . \qquad (4)\end{aligned}$$

As such, the model-based reliability of the scale score will be the ratio of the variance that is explained by the factor ξ to the variance of the scale in total:

$$\begin{aligned}&\text{rel(tau-equivalent scale)}\\&= \frac{\text{variance explained by } \xi}{\text{total scale variance}}\\&= \frac{J^2\lambda^2}{J^2\lambda^2 + \sum_{i=1}^{J}\theta_i} \ . \qquad (5)\end{aligned}$$

This is, in fact, a model-based estimate of Cronbach's α (Cronbach, 1951; Guttman, 1945; Miller, 1995), which makes the assumption of tau-equivalence. Indeed, such an assumption is critical for the derivation of the more familiar closed-form computational formula

$$\alpha = \left(\frac{J}{J-1}\right)\left[1 - \frac{\sum_{i=1}^{J}\sigma_{X_i}^2}{\sigma_{\text{scale}}^2}\right], \qquad (6)$$

where $\sigma_{X_i}^2$ is the i^{th} scale item's variance and σ_{scale}^2 is the variance of the total scale scores. Using the tau-equivalent model parameter estimates from Table 2 within Equation 5 results in a model-based reliability estimate whose pieces are

$J^2\lambda^2 = 6^2(0.910)^2 = 29.812$ and,

$$\sum_{i=1}^{J}\theta_i = (2.353 + 0.645 + 0.553 + 1.260 + 3.376 + 3.862) = 12.049 ,$$

which assemble to yield

$$\hat{\alpha} = 29.812/(29.812 + 12.049) = 0.712$$

It is important to note that this diverges from the reliability estimate derived from the application of Equation 6 to the raw data, which yields $\hat{\alpha}$ = 0.786. The source of this divergence is rooted in the fact that the parameter estimates from the tau-equivalent model used in Equation 5 were derived by forcing loading equivalence while the values used in Equation 6 merely assumed such equivalence but without any formal constraints within an SEM.

So which is correct, the one estimating that 71.2% of the variance of scale scores is attributable to true score variability or the one estimating 78.6%? Quite simply, neither, as the model of tau-equivalence is incorrect and severely so according to the fit indices in Table 2. Hence, estimates of reliability based on Cronbach's α are of limited utility for these data. This would not have been known, however, without assessing model-data fit for the tau-equivalent model, thus further underscoring the value of assessing the fit of the model that is formally underlying one's choice of reliability coefficient.

For the more flexible congeneric model, again using either the linear algebra of composites or path tracing, the total variance in the scale score for a J-item instrument with no error covariances may be shown to be:

$$\begin{aligned}&\text{var(congeneric scale)}\\ &= \text{variance explained by } \xi + \text{variance explained by error terms}\\ &= \sum_{i,j=1}^{J}\lambda_i\lambda_j + \sum_{i=1}^{J}\theta_i = (\sum_{i=1}^{J}\lambda_i)^2 + \sum_{i=1}^{J}\theta_i . \qquad (7)\end{aligned}$$

The resulting model-based reliability is thus:

$$\begin{aligned}&\text{rel(congeneric scale)}\\ &= \frac{\text{variance explained by } \xi}{\text{total scale variance}}\\ &= \frac{(\sum_{i=1}^{J}\lambda_i)^2}{(\sum_{i=1}^{J}\lambda_i)^2 + \sum_{i=1}^{J}\theta_i} . \qquad (8)\end{aligned}$$

This quantity, referred to by McDonald (1999) as ω, makes no assumptions about equality of true score or error score variances although it still assumes zero correlations among errors and of errors with ξ. Using the congeneric model parameter estimates from Table 2 within Equation 8 yields a model-based reliability estimate whose pieces are

$$(\sum_{i=1}^{J}\lambda_i)^2 = (1.581 + 0.856 + 0.825 + 0.413 + 1.725 + 1.208)^2 = (6.608)^2 = 43.666 \text{ and}$$

$$\sum_{i=1}^{J}\theta_i = (1.510 + 0.717 + 0.631 + 1.142 + 2.262 + 3.688) = 9.950, \text{ which assemble to}$$

yield $\hat{\omega} = 43.666/(43.666 + 9.950) = 0.814$.

That is, we estimate that 81.4% of the variance of scale scores is attributable to true score variability, as opposed to lower estimates from both the computational and model-based α estimates we obtained above.

Indeed, α is occasionally described as a lower-bound estimate of scale reliability; in practice, however, α estimates can exceed ω estimates under a variety of circumstances (see, e.g., Dunn, Baguley, & Brunsden, 2014; Widaman, Little, Preacher, & Sawalani, 2011). More important than relative mag-
Nitude, however however, is the issue of accuracy; McDonald's ω has a greater chance of being an accurate representation of scale reliability given its less heavy reliance on unrealistic assumptions.

Whereas in the above example we computed the ω estimate manually from the necessary congeneric model parameter estimates, one may also compute reliability directly within the modeling software. This may be done in a manner that cleverly parallels the models in Figure 5 (see, e.g., Miller, 1995; Raykov, 1997), but such modeling tricks are now largely unnecessary. In M*plus*, for example, an additional parameter may be created to represent ω, which is precisely the function in Equation 8.

At first this may be seen to hold little advantage, given the ease of the hand computation. However, a by-product of estimating the reliability coefficient with the modeling software is the ability to generate an accompanying *confidence interval*. Although the *asymptotic standard error* for the additional parameter (i.e., ω) is typically estimated through what is known as the *delta method* (i.e., a first-order Taylor series based approximation) (see, e.g., Casella & Berger, 2002), which in turn requires the assumption of normality of ω estimates to construct a confidence interval, a potentially more accurate confidence interval may be derived through *bootstrap resampling* (Hancock & Liu, 2012).

Appendix B presents extended M*plus* code for the congeneric model such that both a point estimate and a bias-corrected
bootstrap interval estimate for ω are contained in the output. The latter is based on 5000 bootstrap samples, with inequality constraints to keep loadings from going negative (see Hancock & Nevitt, 1999). The resulting point estimate is the same as computed previously, $\hat{\omega}$=0.814, and the associated 95% bias-corrected bootstrap confidence interval for ω is (0.777, 0.844). The asymmetry of this interval around the point estimate underscores the value of the bootstrap interval relative to one that could be constructed based on the asymptotic standard error accompanying the computation of $\hat{\omega}$.

Implications for Practitioners

Because of its less stringent aassumptions, McDonald's ω is widely viewed within the measurement community as a superior alternative to Cronbach's α because it allows analysts to estimate reliability in a manner that is far more likely to be consistent with a scale's underlying congeneric measurement structure (see, e.g., Dunn et al., 2014; McNeish, in press; Sijtsma, 2009). Indeed, even researchers outside the measurement community have recognized the limitations of α (e.g., Crutzen & Peters, 2017).

Nonetheless, the use of α persists among research practitioners, in part due to its wide-ranging accessibility. Specifically, Cronbach's α has both a model-based formula (Equation 5) and a more common closed-form computational formula (Equation 6), the latter of which is available in common statistical software packages such as *SPSS*, *SAS*, and *Stata*. McDonald's ω, on the other hand, requires the use of a model-based formula although we have recently proposed a closed-form computational formula for ω (Hancock & An, 2016). Until recently, this has meant that, applied researchers have needed to be familiar with SEM packages and conduct analyses like those presented in the last section. Every bit as accurate today is the wise observation by Borsboom (2006) over a decade ago: "there is little chance of convincing [psychologists] to use a model—any model—that is not "clickable" in the menus of major statistical programs" (p. 433).

Fortunately, more recently, several *R* packages have been developed to compute ω directly from a unidimensional confirmatory factor model, including MBESS (Kelley, 2015), semTools (Pornprasertmanit, Miller, Schoemann, & Rosseel, 2013), and `coefficientalpha` (Zhang & Yuan, 2015); some even offer confidence intervals assuming normality of $\hat{\omega}$ estimates. For the simulated dataset, we used each of these packages to estimate ω as well as Cronbach's α for reference. The necessary R code and output appear in Appendix C, with results matching the previous model-based point estimates of $\hat{\omega}$=0.814 and $\hat{\alpha}$=0.786 from M*plus*.

There are, however, limitations of these *R* package that bear making explicit. First, and most obvious, is that the applied researcher is required to know how to use *R*. Although *R*'s popularity is continually growing, it is by no means ubiquitous, and its general lack of graphic interface will likely preclude its universal adoption among practitioners (although *RStudio* and the *Shiny* provide potential pathways for overcoming these limitations).

Second, although various model-assumption diagnostics are available in some packages (e.g., `coefficientalpha`'s *F*-test of tau-equivalence, as seen in Appendix C), these packages are far from exhaustive in their capabilities. In general, full-fledged SEM packages have enhanced diagnostic capabilities and added versatility for accommodating various real data challenges associated with reliability estimation. Therefore, estimating reliability indices with a full-fledged SEM package remains, for now, likely the preferable approach for data analysts, even though it requires a good deal of front-end investment for those unfamiliar with such modeling packages.

Diagnostically, even if one were by-passing parallel and tau-equivalent models in favor of the congeneric model, which seems entirely reasonable for most real data sets, the congeneric model may still not provide satisfactory fit to the data. One reason may be *local dependence*, which can be diagnosed with the judicious use of so-called *modification indices* (see, e.g., Mueller & Hancock, 2010). While not infallible, these indices are able to point toward relations Specifically, the error covariances contribute to the total scale variance in the denominator, as seen in an expansion of Equation 8:

$$\text{rel(congeneric scale)} = \frac{(\sum_{i=1}^{J}\lambda_i)^2}{(\sum_{i=1}^{J}\lambda_i)^2 + \sum_{i=1}^{J}\theta_i + 2\sum_{i>j}^{J}\theta_{ij}} . \qquad (9)$$

Failure to detect and accommodate such error covariances could make the reliability estimate in Equation 8 an over-estimate (if the error covariances are positive) or an under-estimate (if the error covariances are negative). One should be also aware that such error covariances generally signify additional dimensions at work among pairs or subsets of items. Whether their existence warrants modification of the scale items, and/or a complete rethinking of the dimensionality of the instrument at hand, however, remains up to the researcher.

Other reasons data-model fit may be poor when evaluating a measurement model could include failures to meet assumptions underlying the estimation process itself (e.g., maximum likelihood). Fortunately, the SEM framework allows many such violations to be remediated. Nonnormality, for example, can be addressed through, say, bootstrapping or Satorra-Bentler rescaling corrections to parameter test statistics and fit indices (see Finney & DiStefano, 2013). Nonrandom samples, such as those arising through *complex/multilevel sampling structures* (e.g., multistage sampling, sampling weights), can be accommodated through design-based corrections to standard errors and fit statistics (see Stapleton, 2013). The SEM framework is also quite adept at handling *item-level missingness* through, for example, *full information maximum likelihood* estimation or *multiple imputation* (see, e.g., Enders, 2013). Thus, there are many benefits to taking this more comprehensive model-based approach to scale reliability assessment.

Conclusion

As mentioned at the start of this ITEMS module, different areas of quantitative methods have evolved at different rates to meet different, but often related, goals. Reliability assessment, although having evolved primarily in a measurement/ /psychometric arena, is nicely subsumed by the latent variable / SEM framework that has developed in a more applied statistical domain. As we have demonstrated in this ITEMS module, embedding the practice of scale reliability assessment within the SEM framework has benefits in terms of the articulation of the measured and latent variables and their links. Through such articulation, this approach facilitates a formal evaluation of the relevant models underlying the reliability assessment, allows one to model additional relations and address potential data challenges at hand, and, ultimately, yields more appropriate point and interval estimation of reliability.

In this ITEMS module we have also been rather unveiled in our sentiments toward Cronbach's α. As McNeish (in press) nicely put it, "Cronbach's alpha had a good run and was able to hold down the fort for the field for over 50 years, but methodological reinforcements have indeed arrived." The reinforcement we have emphasized is McDonald's ω, which we believe to be a reasonable successor. It is not the only contender, however.

Variations on ω itself exist, for a variety of specific scenarios, include *omega hierarchical* (see Kelley & Pornprasertmanit, 2016; Zinbarg, Revelle, & Yovel, & Li, 2005) and what McNeish referred to as *Revelle's omega total* (Revelle & Zinbarg, 2009). Other options also exist, including, but not limited to, *greatest lower bound reliability* (Jackson & Agunwamba, 1977; Moltner & Revelle, 2015), *maximal reliability* (e.g., Bentler, 2007; Hancock & Mueller, 2001; Raykov, 2004), *Bentler's rho* (Bentler, 1968), the *explained common variance* method (see Sijtsma, 2009), as well as indices such as so-called *construct reliability* and *average variance explained* (Fornell & Larcker, 1981) that attempt to assess the reliability of the underlying construct itself (cf., Hancock & Mueller, 2001).

Whichever one, or ones, wind up succeeding Cronbach's α, more important is that (1) something does, immediately, and (2) that it be approached from within a formal SEM framework. For the uninitiated practitioner, there are clear startup costs to learning the necessary modeling procedures. We believe, however, along with many others (e.g., Crutzen & Peters, 2017; Green & Yang, 2009; Peters, 2014; Schmitt, 1996), that this investment is well worth the effort for the future of scale development and evaluation.

Glossary

True score – the long run expected value of a measure variable (e.g., of a scale item) for a given individual.

Error score – the difference between an individual's observed score on a variable (e.g., on a scale item) and the individual's true score for that item.

Parallel scale items – items on a unidimensional scale in which all items have the same amount of true score variance and error score variance.

Tau-equivalent scale items – items on a unidimensional scale in which all items have the same amount of true score variance, although potentially different error score variance.

Congeneric scale items – items on a unidimensional scale in which all items have potentially different true score variance and potentially different error score variance

Cronbach's α – a measure of scale reliability for unidimensional scales that assumes tau-equivalent scale items.

McDonald's ω – a measure of scale reliability for unidimensional scales that assumes congeneric scale items.

Self-assessment

For this self-assessment, imagine that you have a unidimensional five-item scale, for which a simple-sum total score is desired, similar to the example in the module. The first part of this assessment revolves around the conceptual principles and best practices surrounding the assessment of that total score's reliability. The second part of the assessment is performance based and uses the n = 300 simulated cases from the first five items of the "Task Goal Orientation" scale (i.e., ignoring item 6). These and other questions are also available in the interactive module on the NCME website.

Part 1 – Conceptual Foundations

1) For a five-item scale, how many unique variances and covariances exist among those scale's items?
2) Within an SEM framework, how many parameters would a parallel model, tau-equivalent model, and congeneric model each have for explaining the items' variances and covariances? What would those parameters be, specifically?
3) How many *df* would each of the parallel, tau-equivalent, and congeneric model have?
4) Which of these models would meet the assumptions of Cronbach's α?

Answers to Part 1

1) There are 5 unique item variances and 10 unique item covariances, for a total of 15 unique values in the item covariance matrix.
2) Parallel model: 2 parameters (one common loading λ and one common error variance θ).
   Tau-equivalent model: 6 parameters (one common loading λ and potentially different error variances $\theta_1 \ldots \theta_5$).
   Congeneric model: 10 parameters (potentially different loadings $\lambda_1 \ldots \lambda_5$ and potentially different error variances $\theta_1 \ldots \theta_5$).
3) The parallel model has 15-2=13 *df*; the tau-equivalent model has 15-6=9 *df*; and the congeneric model has 15-10=5 *df*.
4) Both the tau-equivalent and parallel models meet the assumptions of Cronbach's α (with the parallel model being unnecessarily stringent).

Part 2 – Applications

Please use an SEM package of your choice to complete the following exercises.

5) Based on the raw data, use the computational formula in Equation 6 to estimate Cronbach's α. You may do this by getting the necessary descriptive statistics to populate Equation 6 or use the software package of your choice to directly estimate α.
6) Using maximum likelihood estimation obtain data-model fit indices (e.g., χ^2, AIC, BIC, SRMR, RMSEA, CFI) for the parallel, tau-equivalent, and congeneric models (double check that your *df* are correct for each model.)
7) Based on the parameter estimates from the tau-equivalent model, determine the model-based estimate of Cronbach's α using Equation 5.
8) Based on the information criteria, which model would you select? Based on the χ^2 difference tests, which model would you select?
9) Based on the parameter estimates from the congeneric model, use Equation 8 to estimate McDonald's ω.
10) If your SEM software has the capabilities to model ω as an additional parameter, do so and verify that your resulting value matches the one you computed in question 9.
11) Based on the accompanying maximum likelihood asymptotic standard error, compute and interpret a 95% confidence interval for ω.
12) If your SEM software has bootstrapping capabilities, compute and interpret a 95% bootstrap confidence interval (bias-corrected or not) using 5000 bootstrap resamples.

Answers to Part 2

5) Using *SPSS*, the computation estimate using the first five scale items is $\hat{\alpha}$ =.781.
6) Using M*plus*:
   Parallel model: χ^2=408.931, AIC=412.931, BIC=420.339, SRMR=.604, RMSEA=.319, CFI=.153.
   Tau-equivalent model: χ^2=155.954, AIC=167.954, BIC=190.177, SRMR=.300, RMSEA=.233, CFI=.686.
   Congeneric model: χ^2=8.476, AIC=28.476, BIC=65.514, SRMR=.018, RMSEA=.048, CFI=.993.
7) $\hat{\alpha}$ =.704. (Note that this value differs from that computed in Question 5. As mentioned in the article, this is because the model-based value uses loadings forced to be equivalent whereas the traditional value merely assumes such equivalence.)
8) Lowest AIC and BIC values suggest the congeneric model.
   χ^2 difference tests also suggest congeneric to be statistically significantly better than the others (congeneric vs. parallel χ^2_{diff} = 400.455, df=8, p<.001;

congeneric vs. tau-equivalent χ^2_{diff} = 147.478, df=4, p<.001)

9) $\hat{\omega}$=0.825.

10) Yes, the M*plus* estimate matches: $\hat{\omega}$=0.825.

11) The asymptotic standard error for $\hat{\omega}$ is 0.016. The 95% confidence interval for ω around 0.825 is (0.794, 0.856).

12) Using M*plus*, the 95% bias-corrected bootstrap confidence interval is (0.790, 0.853).

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Bentler, P. M. (2007). Covariance structure models for maximal reliability of unit-weighted composites. In S. Lee (Ed.), *Handbook of computing and statistics with applications: Vol. 1*. Handbook of latent variable and related models (pp. 1-19). New York: Elsevier.

Bentler, P. M. (1968). Alpha-maximized factor analysis (Alphamax): Its relation to alpha and canonical factor analysis. *Psychometrika*, *33*, 335-345.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425-440.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Wadsworth.

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

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

Crutzen, R., & Peters, G. J. Y. (2017). Scale quality: Alpha is an inadequate estimate and factor-analytic evidence is needed first of all. *Health Psychology Review*, *11*, 242-247.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*, 399-412.

Enders, C. K. (2013). Analyzing structural equation models with missing data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 493-519). Charlotte, NC: Information Age Publishing.

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 439-492). Charlotte, NC: Information Age Publishing.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*, 39-50.

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

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*, 255-282.

Hancock, G. R., & An, J. (2016, April). *Closed-form alternatives for estimating omega reliability*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Hancock, G. R., & Liu, M. (2012). Bootstrapping standard errors and data-model fit statistics. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 296-306). New York: Guilford Press.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future—A festschrift in honor of Karl Jöreskog* (pp. 195-216). Lincolnwood, IL: Scientific Software International.

Hancock, G. R. & Nevitt, J. (1999). Bootstrapping and identification of exogenous latent variables within structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 394-399.

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. *Psychometrika*, *42*, 567-578.

Kelley, K. (2015). MBESS (Version 4.0.0 and higher) [computer software and manual]. Accessible from http://cran.r-project.org.

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods*, *21*, 69-92.

Kline, R. (2016, 4th ed.). *Principles and practice of structural equation modeling*. New York, NY: Guilford.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Erlbaum.

McNeish, D. M. (in press). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*.

Midgley, C., Kaplan, A., Middleton, M., Maehr, M. L., Urdan, T., Anderman, L. H., Anderman, E., & Roeser, R. (1998). The development and validation of scales assessing students' achievement goal orientations. *Contemporary Educational Psychology*, *23*, 113-131.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*, 255-273.

Moltner, A., & Revelle, W. (2015). *Find the greatest lower bound to reliability*. Available online at: http://personality-project.org/r/psych/help/glb.algebraic.html

Mueller, R. O., & Hancock, G. R. (2010). Structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 371-383) New York: Routledge.

Peters, G. J. Y. (2014). 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. *European Health Psychologist*, *16*, 56-69.

Pornprasertmanit, S., Miller, P., Schoemann, A., & Rosseel, Y. (2013). semTools: Useful tools for structural equation modeling. *R Package available on CRAN*.

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

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy*, *35*, 299-331.

Reuterberg, S., & Gustafsson, J. (1992). Confirmatory factor analysis and reliability: Testing measurement model assumptions. *Educational and Psychological Measurement*, *52*, 795-811.

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

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350-353.

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

Stapleton, L. M. (2013). Multilevel structural equation modeling with complex sample data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 521-562). Charlotte, NC: Information Age Publishing.

Traub, R. E., & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice*, *10*, 37-45.

Widaman, K. F., Little, T. D., Preacher, K. J., & Sawalani, G. M. (2011). On creating and using short forms of scales in secondary research. In K. H. Trzesniewski, M. B. Donnellan, & R. E. Lucas (Eds.), *Secondary data analysis: An introduction for psychologists* (pp. 39-61). Washington, DC: American Psychological Association.

Zhang, Z., & Yuan, K. H. (2015). Package 'coefficientalpha'. *R Package available on CRAN*.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 123-133.