Journal of Modern Applied Statistical Methods

Volume 17 | Issue 1 Article 15

6-26-2018

Internal Consistency Reliability in Measurement: Aggregate and Multilevel Approaches

Georgios Sideridis
Harvard Medical School, georgios.sideridis@childrens.harvard.edu

Abdullah Saddaawi
King Saud University, alsadaawi@gmail.com

Khaleel Al-Harbi
Taibah University, k.harbi@qiyas.org

Internal Consistency Reliability in Measurement: Aggregate and Multilevel Approaches

Cover Page Footnote
Address correspondence to Georgios D. Sideridis, Ph.D., Boston Children's Hospital, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115. Email: Georgios.sideridisi@childrens.harvard.edu or Georgios.sideridis@gmail.com.

Journal of Modern Applied Statistical Methods
May 2018, Vol. 17, No. 1, eP2560
doi: 10.22237/jmasm/1530027194

בס"ד

ISSN 1538 − 9472

EMERGING SCHOLARS

Internal Consistency Reliability in Measurement: Aggregate and Multilevel Approaches

Georgios Sideridis
Harvard Medical School
Boston, MA

Abdullah Saddaawi
King Saud University
Riyadh, Saudi Arabia

Khaleel Al-Harbi
Taibah University
Medina, Saudi Arabia

The purpose of this study was to evaluate the internal consistency reliability of the General Teacher Test assuming clustered and non-clustered data using commercial software (Mplus). Participants were 2,000 testees who were selected using random sampling from a larger pool of examinees (more than 65k). The measure involved four factors, namely: (a) planning for learning, (b) promoting learning, (c) supporting learning, and (d) professional responsibilities, and was hypothesized to comprise a unidimensional instrument assessing generalized skills and competencies. Intra-class correlation coefficients and variance ratio statistics suggested the need to incorporate a clustering variable (i.e., university) when evaluating the factor structure of the measure. Results indicated that single level reliability estimation significantly overestimated the reliability observed across persons and underestimated the reliability at the clustering variable (university). One level reliability was also, at times, lower than the lowest acceptable levels leading to a conclusion of unreliability whereas multilevel reliability was low at the between person level but excellent at the between university level. It was concluded ignoring nesting is associated with distorted and erroneous estimates of internal consistency reliability of an ability measure and the use of MCFA is imperative to account for dependencies between levels of analyses.

Keywords: Internal consistency reliability, multilevel structural equation modeling, tau equivalence, Guttman's lambda coefficients

Introduction

Inconsistent measurement is undoubtedly one of the biggest threats to the internal validity of studies and has attracted the interest of researchers since early 1900

doi: 10.22237/jmasm/1530027194 | Accepted: May 15, 2017; Published: June 26, 2018.
Correspondence: Georgios Sideridis, georgios.sideridis@childrens.harvard.edu

(Spearman, 1904, 1910). For that purpose, several indices of internal consistency reliability have been developed to properly capture this important measurement characteristic (Anderson & Gerbing, 1982). In the relevant literature there have been diverse opinions regarding internal consistency and reliability with several authors pointing to diverse operational definitions (Hattie, 1985). In the present study the term internal consistency reliability is used and relates to the earlier use of internal consistency. It represents a domain-sampling approach, as true reliability should involve the presence of two measurement points (Guttman, 1945). Amongst indices, Cronbach's alpha coefficient (Cronbach, 1951) has been one of the most widely used indices with more than 250,000 hits in Google's Scholar database; see also Hogan, Benjamin, and Brezinski (2000). This is particularly interesting despite noticeable shortcomings and challenges regarding computation and interpretation (Boyle, 1991; Cortina, 1993; Hayashi, & Kamata, 2005; Henson, 2001; Kopalle, 1997; Liu, Wu, & Zumbo, 2010; Raykov, 2001; Shevlin, Miles, Davies, & Walker, 2000; Streiner, 2003a). Other commonly-used indices involve omega reliability (Raykov, 1997) and maximal reliability *H* (Li, 1997). The purpose of this study is to illustrate, using an example from a National Examination, the measurement of internal consistency reliability under the lenses of multilevel modeling as a means to properly assess the amount of error that the latent trait contains across different levels in the analysis. Here, internal consistency reliability refers to true-score variance. A secondary purpose is to illustrate the estimation of various indices using widely known software.

Inevitably, when assessing internal consistency of a measure one must consider the context in which individuals are located. For example, when students take a test, their scores and performance may be more similar to students within their class compared to students from other classes, schools, or neighborhoods (Opdenakker & Van Damme, 2000). This apparent relationship will likely be reflected with a correlational structure that will account for those dependencies (e.g., an autocorrelation structure if students are nested within time). In the above example, student scores will likely be more strongly correlated when tested within their class (and the mean of their class), compared to across classes (and the grand mean). Ignoring that dependency will likely result in estimates of internal consistency that confound true within and between estimates of reliability as the aggregate term will ignore the true score and error variance estimates at each level, placing them all under a single residual term (Geldof, Preacher, & Zyphur, 2014). As a problem in educational research, it was first described by Robinson (1950) as the ecological fallacy phenomenon, which refers to the implicit assumption that estimates at one level generalize to another (Morin, Marsh, Nagengast, & Scalas,

2014) with several applications supporting opposite claims (e.g., Marsh, 2007; Schwartz, 1994). The basic justification for estimating multilevel reliability is that true score variance may be "captured to a different degree at each level" (Geldof et al., 2014, p. 75).

Within the Structural Equation Modeling (SEM) framework, when ignoring nesting, a factor loading reflects the expected value of change in an indicator when the factor changes by one standardized unit (Pornprasertmanit, Lee, & Preacher, 2014). This relationship, however, between an item and a latent factor, should not presumably be the same if the unit of analysis is the person in relation to his/her cluster's mean (e.g., when students are nested within their class – as in group-mean centering) compared to the person being seen in relation to the whole group (aggregate data, ignoring nesting – as in grand mean centering). The next section describes the estimation of various internal consistency reliability indices.

Internal Consistency Reliability Estimation

Cronbach's Alpha

Based on Classical Test Theory (Nunnally, 1978), a measured item/construct's X score is comprised of two components: a true score T plus some form of error e (i.e., $X = T + e$), with the expectation that error is random rather than systematic. Since we rarely measure single item constructs, unidimensionally-measured phenomena are often described with a single factor model in which items contribute stochastic and white noise information. Using a three-item instrument the 1-factor model is expressed as follows (Wang & Wang, 2012):

$$\begin{aligned} Y_1 &= \lambda_1\xi_1 + \delta_1 \quad (\text{Item } 1) \\ Y_2 &= \lambda_2\xi_1 + \delta_2 \quad (\text{Item } 2) \\ Y_3 &= \lambda_3\xi_1 + \delta_3 \quad (\text{Item } 3) \end{aligned}$$

with each of items Y_1, Y_2, and Y_3 being linked to the latent structure ξ_1 stochastically (with λ being the correlation between the item and the latent dimension) and δ a form of random error as the items are likely imperfect estimates of the true trait. Wang and Wang (2012) emphasized that no matter how carefully the procedures have been implemented or how refined a measure is, error of measurement is sizable and hopefully reflects random rather than systematic variations; for an excellent discussion see Streiner (2003b) and Judd, Smith and Kidder (1991). Based

on the above single factor model and earlier work (Guttman, 1945), Cronbach proposed the alpha statistic as a measure of internal consistency assuming that all items contribute to the measurement of a construct and that contribution is reflected in the intercorrelations between items (i.e., $k * \bar{r}$) as follows:

$$\text{Cronbach } \alpha = \frac{k\bar{r}_i}{1+(k-1)\bar{r}_i} \quad (1)$$

Thus, the term $\bar{r}_i$ reflects the mean intercorrelation between items i_1, i_2,…, i_k and k is the number of items. The above formula was used for presentation only as it is the easiest to conceptualize; for estimation we employed the alternative formula, which has wider use:

$$\text{Cronbach } \alpha = \frac{n^2\sigma_{ij}}{\sigma_{\chi}^2} \quad (2)$$

with n representing the number of items, σ_{ij} the average covariance between items, and σ_{χ}^2 the sum of all item variances plus 2 times the sum of the covariances between items (i.e., scale's variance, see Geldof et al., 2014). Later Cronbach corrected the positive bias that the number of items exerts on the coefficient by adopting the Spearman-Brown formula (J. Brown, 1996) and proposed alternative formulations (Cronbach & Gleser, 1964). Obviously the magnitude of the interitem correlation and the number of items are positive contributors to alpha with larger correlations and lengthy instruments being associated with higher estimates of internal consistency reliability. As several researchers noted, however, Cronbach's coefficient alpha is a lower bound to the true reliability when items are tau equivalent (Lord & Novick, 1968). Consequently, it may seriously underestimate the internal consistency of a measure (Osburn, 2000; Thompson, Green, & Yang, 2010). For that purpose, an alternative to the original formula was implemented which involved a correction for sample size (Kristof, 1963):

$$\text{Cronbach's alpha using Kristof's correction } \alpha_{\text{K}} = \frac{2}{N-1} + \frac{N-3}{N-1}\alpha \quad (3)$$

with N being the sample size.

Cronbach's alpha requires several conditions to be met before its estimates are valid, some of which have been ignored in the literature, to say the least. The most important ones are: (a) item scores should be on an interval level data with no restriction of range (Fife, Mendoza, & Terry, 2012) without having to implement the K-R 20 formula, (b) linearity and homoscedasticity of errors, (c) small amounts of measurement error and correction for attenuation of both variances and covariances, (d) same distributions between items, (e) unidimensionality, (f) absence of systematic sources of error, (g) independence of items in terms of content, (h) tau equivalence (i.e. presence of equal factor loadings across indicators), albeit the fact that the presence of congeneric measures is likely the norm (i.e., different relationships between items and latent variable are observed and different variances of their errors, T. Brown, 2015), and, last, (i) parallel equivalence, a more strict form of tau, in that both the factor loadings and the error variances of the items are considered equal (In the present study both tau equivalent and essentially tau equivalent, i.e. differing from tau equivalent in the presence of an additive constant, measures will be considered as tau equivalent.). Research has shown significant deviations between true and observed point estimates of internal consistency using Cronbach's alpha when its assumptions are not met (Raykov, 2001, 2012), thus questioning its utility under several conditions.

Composite Reliability Omega

Omega reliability (McDonald, 1970, 1999; Raykov, 1997), despite its similarity to alpha, possesses the advantage of allowing for heterogeneous item-latent variable correlations. It is estimated as follows:

$$\text{Omega} = \frac{\left(\sum_i \lambda_i\right)^2}{\left(\sum_i \lambda_i\right)^2 + \sum_i \text{Var}\left(\varepsilon_i\right)} \tag{4}$$

With λ_i being the factor loadings of item i and $\Sigma_i \text{Var}(\varepsilon_i)$ the respective error variances of item i. This formula ignores the likelihood that a correlated structure in the residuals is present, in which case reliability needs to be adjusted accordingly (Westfall, Henning, & Howell, 2012). In instances that correlated errors reflect measurement artifacts (Wang & Wang, 2012) such as presence of a single stem across all items (e.g., "It is important to me to…") or the presence of a third latent aptitude trait (e.g., language, complex terminology) that is a prerequisite to

comprehending the content of some items, one needs to adjust the coefficient for collinearity in the residuals as following:

$$\text{Omega} = \frac{\left(\sum_i \lambda_i\right)^2}{\left(\sum_i \lambda_i\right)^2 + \sum_i \text{Var}\left(\varepsilon_i\right) + 2\sum_i \sum_j \text{Var}_{ij}} \quad (5)$$

With the term $2\Sigma_i\Sigma_j \text{Var}_{ij}$ being two times the sum of the covariance between the error terms, representing a scale's variance. More recently, Zinbarg, Revelle, Yovel, and Li (2005) provided an extension of omega through estimating a lower bound estimate of internal consistency reliability. However, in the present study the intercorrelations of residual estimates were negligible around zero and, thus, this formula was not implemented further.

Maximal Reliability *H*

The *H* coefficient termed maximal reliability *H* (Bentler, 2007) was assessed as a means of estimating reliability using an optimally weighted composite using the standardized factor loadings as follows (Li, 1997; Raykov, 2004):

$$H = \frac{\sum_i \frac{\lambda_i^2}{1-\lambda_i^2}}{1+\sum_i \frac{\lambda_i^2}{1-\lambda_i^2}} \quad (6)$$

With l_i^2 being the standardized factor loading of item *i* squared (Hancock & Mueller, 2001). The advantage of the maximal reliability coefficient compared to omega lies in the fact that negative factor loadings now offer meaningful variance that is modeled properly. Also, the *H* statistic uses a weighted estimate by squaring the individual factor loadings (Hancock & Mueller, 2001) and the estimated reliability can never be less than reliability of the best measured item. Last, the weighing procedure saliently downgrades less informative items which load weakly on the factor (Geldof et al., 2014).

Other Lower-Bound Indices of Reliability

Several reliability coefficients have been developed as lower-bound estimates of true reliability due to the apparent bias observed with methods proposed earlier

(e.g., alpha, Feldt, 2002). The idea that governs those indices is that the covariances between items represent true information whereas the variances of the items contain both true and unique variability. The interested reader can consult the works of Jackson and Agunwamba (1977) for excellent reviews. One such index is α_{pc} (ten Berge & Hofstee, 1999), which employs the eigenvalue of the first principal component, in the case of unidimensional structures. The coefficient is estimated as follows:

$$\alpha_{pc} = \frac{n}{n-1}\left(1-\frac{1}{\lambda_1}\right) \tag{7}$$

With λ_1 being the eigenvalue of the first principal component from a PCA analysis using commercial software; see also Raykov and Pohl (2012) for an alternative conceptualization through modeling common factor variance.

A second index is Guttman's Lambda 1 coefficient. It is estimated as the ratio of one minus the sum of the items' error variances to the instrument's total variance σ_x^2:

$$\lambda_1 = \frac{1-\sigma_{ii}^2}{\sigma_x^2} \tag{8}$$

with the term σ_{ii} representing the sum of the item error variances and σ_x^2 the sum of all item variances plus 2 times the sum of the covariances between items. The idea behind the coefficient is that all of items' information represents measurement error except the interitem covariances, which reflect true variance.

A last index is Guttman's Lambda 2 coefficient which equals lambda 1 when the items are tau equivalent. It is estimated through adding lambda 1 to the ratio of the square root of two times the item covariances squared times n / (n – 1), and all that divided by the sum of all item variances plus 2 times the sum of the covariances between items (i.e., the scale's total variance estimate):

$$\lambda_2 = \lambda_1 = \frac{\sqrt{\frac{n}{n-1}C}}{\sigma_x^2} \tag{9}$$

with *C* being the square root of the sums of squares of the off diagonal elements. It is considered an improved estimate over Cronbach's alpha (Osburn, 2000) and is very similar to the u_2 estimate of ten Berge and Zegers (1978).

Confidence Intervals of Internal Consistency Estimates

Undoubtedly, more attention has been given in the literature on internal consistency reliability point estimation compared to confidence interval estimation (Muthén, 1991; Raykov, 1998, 2002, 2006). Two prominent methods described in the literature involve parametric bootstrapping (Goldstein, 2003; Kuk, 1995) or the delta method (Raykov, 2002) through deriving standard errors with the first-order delta procedure, with a more general method involving bootstrapping percentile confidence intervals (Raykov, 1998; Raykov & Shrout, 2002). Researchers have also employed various software such as Mplus and R (e.g., Dunn, Baguley, & Brunsden, 2013). More recently, Raykov and Marcoulides (2012) introduced the non-bootstrap method with the use of a maximum likelihood estimator (MLR); see also Raykov and Marcoulides (2011). In the present study the estimation of empirically derived asymmetric confidence intervals was implemented in light of the fact that (a) estimated standard errors may be less informative and (b) the distribution of omega, alpha, *H* reliability, or the other coefficients is not known (Raykov, 1998, 2002). Thus, confidence intervals were estimated using the logit transformation in order to normalize the internal consistency estimates with the confidence intervals being estimated using $\hat{z}$ following the lead of Padilla and Divers (2013) and the earlier findings of Raykov (2002). Initially omega or other reliability indices are transformed onto a normal deviate estimate $\hat{z}$ in order to estimate a confidence interval of the form (Raykov et al., 2016; Raykov, Rodenberg, & Narayanan, 2015):

$$\hat{z} \pm z_{\frac{a}{2}} \text{S.E.}\left(\hat{z}\right) \tag{10}$$

With $z_{a/2}$ being the two-sided level of significance for a given alpha level. The logit transformation of omega is given by:

$$\hat{z} = \ln\left(\frac{\hat{\omega}}{1-\hat{\omega}}\right) \tag{11}$$

Table 1. Indices of internal consistency reliability for the aggregate scale and its bifurcation to within and between levels

Level of analysis	Cronbach's alpha	Alpha K-corrected	Omega CR	Guttman's λ_1	Guttman's λ_2	α_{pc}[†]	Maximal *H*
Aggregate scale	0.689	0.692	0.693	0.516	0.69	0.537	0.745
(Single level)	(0.668-0.711)	(0.671-0.714)	(0.672-0.714)	(0.501-0.532)	(0.669-0.712)		(0.726-0.765)
Within level	0.673	0.674	0.677	0.505	0.675	0.527	0.731
(Person)	(0.631-0.718)	(0.632-0.719)	(0.648-0.707)	(0.475-0.537)	(0.633-0.720)		(709-0.755)
Between level	0.981	0.981	0.969	0.735	0.987	0.744	0.97
(University)	(0.431-10.00)	(0.431-10.00)	(0.950-0.989)	(0.322-10.00)	(0.432-10.00)		(0.951-0.990)

Note: Estimates in the parentheses are 95% confidence intervals based on the logit transformation (Raykov, Marcoulides, & Akaeze, 2016).
[†] A standard error for this estimate could not be computed because it was based on an estimate of an eigenvalue for which an error term was not available. Without knowing the distribution of eigenvalues we decided not to attempt to estimate the error terms around those estimates for both the within and between levels in the analysis.
Alpha K-corrected involves Kristof's correction for sample size.
Estimates of Guttman's lambda 1 and lambda 2 coefficients were cross-validated using other commercial software for which routines were readily available.
Upper bound estimates were constrained to unity, when exceeding the theoretical min-max, for ease of interpretation.

and its estimate of standard error:

$$\text{S.E.}\left(\hat{z}\right) = \frac{\text{S.E.}\left(\hat{\omega}\right)}{\hat{\omega}\left(1-\hat{\omega}\right)} \tag{12}$$

These results are shown in Table 1.

Importance of the Present Study

Interestingly, the above mentioned approaches to internal consistency estimation have been primarily implemented ignoring the presence of nested structures. In all these instances, however, within and between level reliability have been confounded as a single estimate which literally reflects an average of the two estimations, thus, conflating the estimates at each level (Geldof et al., 2014; Heck, 1999). Inevitably, one can end up having proper levels of internal consistency at one level in the analysis but low reliability at another level, affecting any subsequent structural relations in unknown ways (Pornprasertmanit et al., 2014). Proper evaluation of within and between level internal consistency will allow the evaluation of subsequent multilevel hypotheses and predictions after evaluating true score estimates at each level in the analysis. Within this notion, any estimate of internal consistency using aggregate data will likely misrepresent the true reliability of a measure in cases where measurement error is markedly different at the within versus the between level of the analysis. Pornprasertmanit et al. (2014) have shown that the aggregate approach may provide unbiased estimates in the case of extremely large ICCs, e.g., > 0.75, which as the authors mentioned is an unrealistic estimate in applied settings, thus recommending the multilevel modeling approach.

It is only when the estimates in each level are identical that the aggregate internal consistency estimation would reflect the true estimate. The problem under study has been illustrated in the findings of G. Woodhouse, Yang, Goldstein, and Rasbash (1996), who demonstrated that after adjusting for the measurement error at the within level (student) slopes at the structural level increased by a factor of 1.27 (for the relationship between year 3 and year 5 performance). Thus, the need to estimate and/or adjust for the measurement error that is present at each level in the analysis has significant implications for evaluating the behavior of predictors; for an applied example see Morin et al. (2014). On the other hand, ignoring correlated structures (nested data) will likely be linked to erroneous estimations of

power and sample sizes as the magnitude of the correlated structure (intraclass correlation) will not be accounted for. In the presence of unreliable measurement and large standard errors, the needed sample sizes will likely be prohibitive. The purpose of the present study was to evaluate and illustrate, using a commercially available software (Mplus Version 7.4), estimation of within and between level internal consistency estimation using a National General Teacher test in Saudi Arabia using the methodologies outlined above for the measurement of various internal consistency reliability indices and using unilevel and multilevel structures. Prior attempts to capture multilevel internal consistency estimation were conducted to evaluate the consistency of means within classes (Raykov & Penev, 2010) rather than the internal consistency of scales; see also Wilhelm and Schoebi (2007) and Huang and Weng (2012) for alternative conceptualizations.

Method

Participants and Measure

Participants were a random sample of 2,000 individuals who had taken the General Teacher Test from the National Center for Assessment in Higher Education. The purpose of this test is to ensure that teachers possess the minimum qualifications required by the state to obtain teaching positions on various disciplines. The general teacher test is comprised of 26 subject-specific subject tests for further specialization. The mean age of the participants was 26.82 years (S.D. = 4.79 years). There were 635 males (31.8%) and 1365 females (68.3%). Participants came from 23 higher education establishments with the number of applicants per institution ranging between 22 and 202 participants. The measure includes four constructs: (a) *planning for learning* to ensure basic literacy and numeric skills as well as a deep understanding of the learning process, (b) *promoting learning*, which evaluates teaching strategies, (c) *supporting learning*, which tests teachers' capacity to establish a safe and conducive to learning environment, and (d) *professional responsibilities*, which evaluates professionalism and self-reflection. The four factors were correlated significantly with each other with Pearson estimates ranging between 0.280 and 0.497, all being significant at $p < 0.001$ (in light of the relatively large sample size). For the purposes of the present study the four subconstructs were considered items that defined a unidimensional ability structure so that the software will be programmed without the added complexity of a lengthy measure.

Data Analysis

Data were analyzed using Mplus (see supplemental content) and modeled the above indices of internal consistency reliability for unilevel (aggregate) and multilevel data (Muthén, 1989, 1990, 1994; Yuan & Bentler, 2007). The university comprised the clustering variable, with students nested within universities, as it would be important to test how reliable an aptitude test's scores are at both the person and the university levels provided the medium of education is through universities. Thus, assessing ability at the university level will allow proper evaluation and comparison between universities and their respective departments, knowing that rankings and ratings are oftentimes conducted at the university and/or the department level. Federal and state agencies may make use of such data. For example, K. Woodhouse (2015) reported that most of governmental funding in 2013 was directed to community colleges and small universities, leaving research institutions to seek funding from other sources. Such decisions need to be granted on hard evidence relating the qualities of universities and departments, thus, accurate estimation of these attributes that the level is essential. Consequently, the estimation of internal consistency reliability at the person level would suggest the precision in which ability can be estimated for each individual (person level), whereas estimation of internal consistency reliability at the university level would reflect aggregate estimates for groups of individuals who belong to a university and would point to the accuracy of estimating aptitude at the organization level (university), so that direct comparisons across universities can be accomplished.

Multilevel Structural Equation Modeling (MSEM) evaluates measurement and structural models at more than one level in the analysis when nesting is in place (Geldof et al., 2014; Heck & Thomas, 2015). The primary purpose of modeling data at two or more levels is to avoid the violation of the independence of observation assumption which is introduced when ignoring the clustering effects (e.g., the effects a school administration, teacher, school culture, or classroom climate exerts on all students-causing a baseline between person correlation that reflects a systematic source of measurement error) (Julian, 2001). Some background information and a description of the models utilized in the present study are presented below.

For the measurement model and using a unidimensional structure, the relations between the items and the latent factor can be expressed using the following equation using scalar expression:

$$Y_i = \mathbf{\Lambda} * \eta_i + \varepsilon \tag{13}$$

With Y_i being the observed items for person i, $\mathbf{\Lambda}$ being the matrix of factor loadings on latent variables η_i, and ε the error terms. In order to accommodate predictors at the latent variable level, structural models can be expressed as follows:

$$\eta = \alpha + \mathbf{B}\eta + \zeta \tag{14}$$

or in the following matrix form

$$\begin{aligned} \eta &= \alpha + \mathbf{B} * \eta + \Gamma * X + \zeta \\ [\eta] &= [\alpha] + [B_{11} \quad B_{12} \quad B_{13} \quad B_{14}][\eta] + [\gamma_1][\text{predictor}] + [\zeta] \end{aligned} \tag{15}$$

With both errors of the measurement and structural models being distributed approximately multivariate normal:

$$\begin{aligned} \varepsilon_j &= \text{MVN}(0, \Theta) \\ \zeta_j &= \text{MVN}(0, \Psi) \end{aligned}$$

Using the multilevel confirmatory factor analysis framework and matrix notation, the following model was fit to the data:

$$Y_{ij} = \mathbf{\Lambda}\eta_{ij} \tag{16}$$

with the general achievement y of student i in university j being a function of vectors of regression coefficients $\mathbf{\Lambda}$ and random effects $\boldsymbol{\eta}$. Thus, what is added in the multilevel framework is the subscript j to indicate that the respective estimates vary across clusters, i.e. universities in the present context.

Results

Prerequisite Analyses

Initially, the necessity to model the university as a random effect was tested by inspecting the Intra Class Correlations (ICCs), the variance ratio statistic, along with the design effect estimate (see Table 2, Kish, 1965; Preacher & Selig, 2012; Werts, Linn, & Jöreskog, 1974). Results justified the presence of multilevel modeling (nonzero ICCs and large between to within variance ratio statistics as

well; > 2 design effect values). Furthermore, the confidence intervals of those estimates did not contain zero suggesting the absence of negligible effects.

Table 2. Intra-class correlation coefficients and variance ratio statistics of the general test with 95% confidence interval estimates

Construct	ICC	ICC-CI	Variance ratio test†	Ratio test CI	DEFF
Planning for Learning	5.40%	2.7-10.7%	33.60%	17.3-64.9%	5.665
Promoting Learning	4.30%	2.0-9.33%	32.00%	20.7-49.4%	4.713
Supporting Learning	3.60%	1.5-8.6%	32.50%	20.7-51.1%	4.105
Professional Responsibilities	1.30%	0.3-4.3%	25.30%	14.3-44.9%	2.138

Note: Each of the four factors represents an item for the purposes of the present study.
† Refers to the Raykov et al. (2016) recent Ratio statistic of the between to within variance estimate as a supplement to the ICC. Confidence intervals are at 95%.
The DEFF estimate refers to the design effect for which values greater than 2.0 suggest that the clustering variable contains information that need to be modeled via multilevel modeling techniques (Maas & Hox, 2005; Muthén & Satorra, 1995). It is estimated using the formula Design Effect = 1 + (Average Cluster Size – 1) * ICC.
Confidence intervals of the ICCs were estimated using the ci.r function from Raykov and Marcoulides (2012). A slightly improved function has been put forth by Raykov et al. (2016).
The ICC was measured as the ratio of the between level variance to the sum of between and within variance: $s_b^2 \Big/ \left(s_b^2 + s_w^2\right)$.

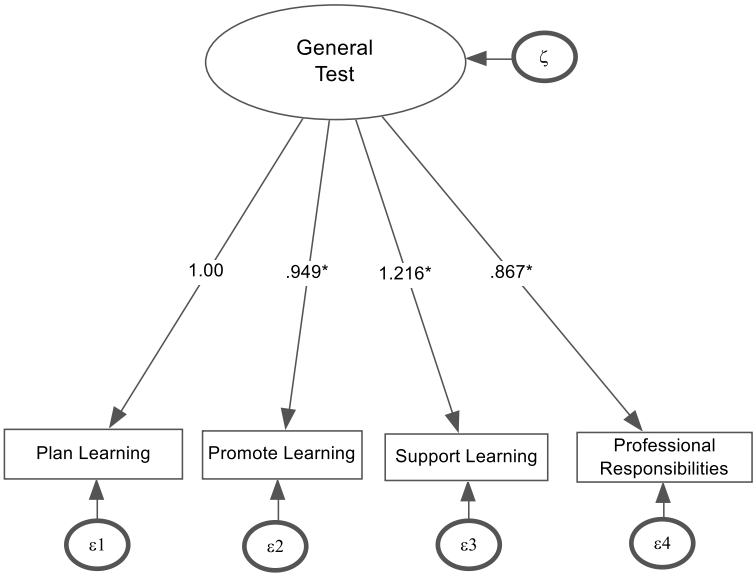

Figure 1. One-factor model for the measurement of general ability using aggregate scores (sum of items) from each of four general ability factors; unstandardized factor loadings are shown

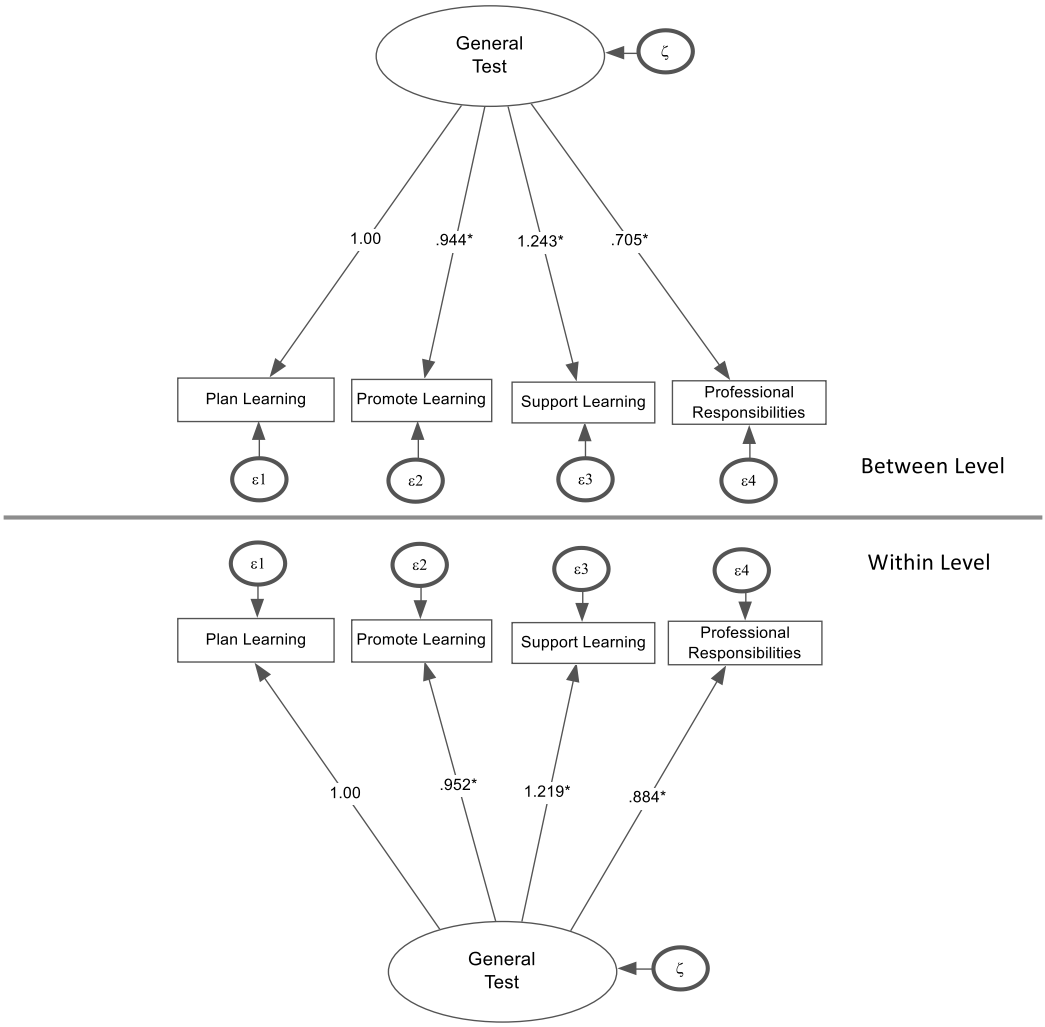

Figure 2. One-factor model at both levels in the analysis (person and university) using unstandardized estimates

Furthermore, the factor structure and consistency of the measure were evaluated using CFA analysis using both single level data and through including nesting due to the university the examinees attended. With the aggregate data, results indicated that the data fit this model well, which run with only 2 dfs, with all items (aggregates) being significant in defining general ability [$\chi^2(2) = 1.530$, $p = 0.465$; RMSEA < 0.001, SRMR = 0.005, CFI = 1.00, TLI = 1.00]. With the clustered data, results pointed to again good model fit with the omnibus chi-square test being non-significant [$\chi^2(8) = 22.096$, $p = 0.005$]. Also descriptive fit indices

along with unstandardized residuals were excellent [RMSEA = 0.030, SRMR = 0.006/0.0079 for both within (person) and between (university) levels, respectively, CFI = 0.995, TLI = 0.992] suggesting a properly measured univariate construct. A last prerequisite analysis involved testing tau equivalence (Graham, 2006; Raykov, 1997), which posits that items contribute equally to the measurement of a latent trait. A two-step approach was followed: First, the equivalence of factor loadings was tested using the aggregate data followed by the equivalence between factor loadings across level of the analysis (i.e., which is a measure of metric invariance rather than a test of tau equivalence). Results using the unilevel approach indicated that constraining all items to contribute equally to the measure of general ability (see estimates in Figure 1) was associated with significantly inferior fit [$\Delta\chi^2(3) = 32.712$, $p < 0.001$], compared to freely estimating those factor loadings, pointing to the absence of tau equivalence.

Using a one-factor simple structure at both levels in the analysis (MSEM) results after constraining the factor loadings to be equivalent in each level suggested the absence of metric invariance [$\Delta\chi^2(4) = 27.750$, $p < 0.001$]. Thus, the measurement of general ability was congeneric and variable at each level in the analysis (see Figure 2). These findings from the aggregate analysis certainly discourage the use of Cronbach's alpha, which will be nevertheless estimated due to frequency of its use and familiarity of the research community with its estimation.

Single Level Reliability Estimates and their Confidence Intervals (Aggregate Data)

When ignoring the nesting structure due to different institutions, results indicated that Cronbach's alpha was equal to 0.689 [C.I. = 0.671-0.706], which is not acceptable, except its upper confidence interval limit which was rather borderline. After applying Kristof's correction (Kristof, 1963), the coefficient was slightly improved with a point estimate equal to 0.692 (compared to 0.689) and 95% confidence intervals of [C.I. = 0.674-0.709]. The Omega index of reliability was equal to 0.693 [C.I. = 0.674-0.711] and maximal reliability equal to 0.745 [C.I. = 0.730-0.763]. α_{pc} was equal to 0.537 and lambda 1 and lambda 2 coefficients were 0.516 and 0.690, respectively [lambda 1 C.I. = 0.503-0.529; lambda 2 C.I. = 0.673-0.708]. Those estimates should be viewed under the lenses of the multilevel structure and the estimates observed at the between person and between university level as shown below before concluding on the adequacy of internal consistency reliability.

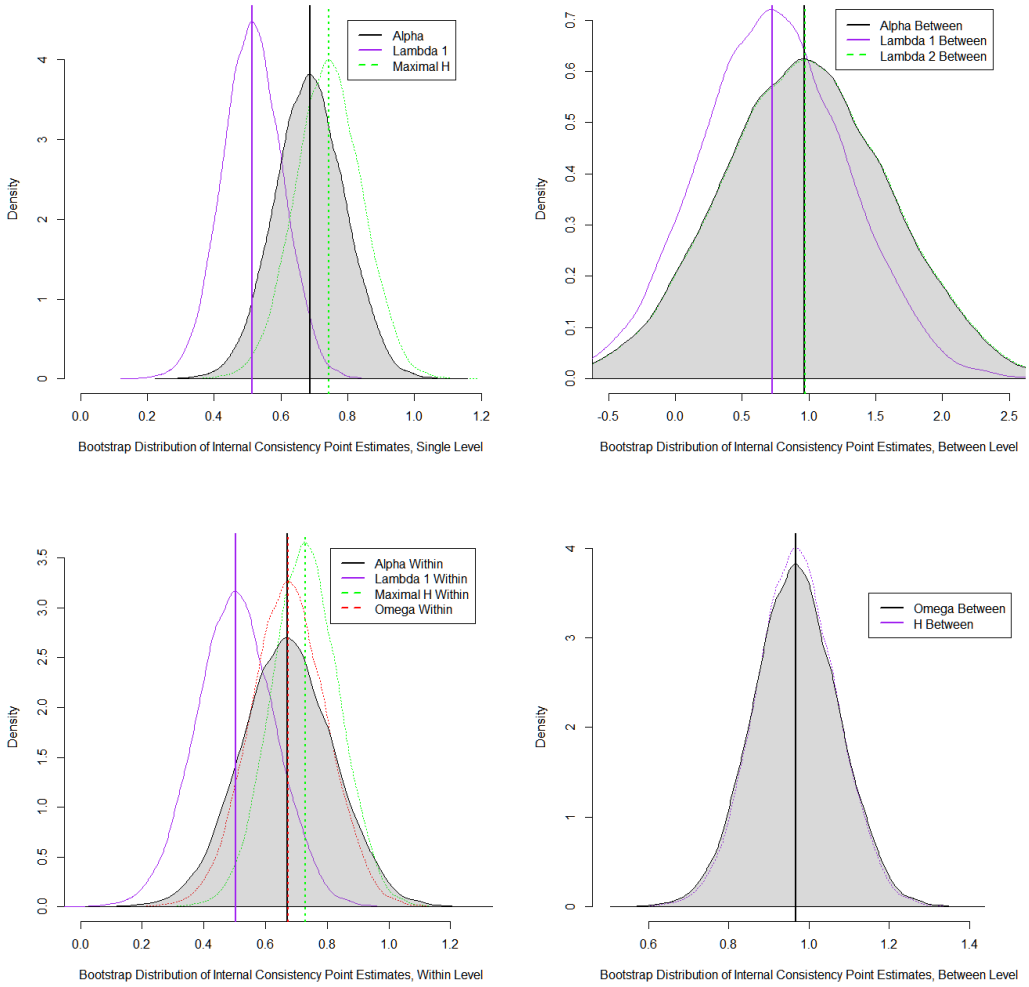

Figure 3. Distributions of internal consistency estimates at the within and between levels in the analysis; coefficients that overlap are now shown; as shown above, the distribution of those estimates is approximately normal, thus, there was no need to apply a log-odds transformation prior to the simulation

---

Reliability Analysis of General Ability Measure at Both the Within and Between Levels of the Analysis (Multilevel Data)

Results with regard to the internal consistency estimates presented above suggested salient differences between the estimates obtained at both the within and between levels in the analysis. The point estimates of the two reliability coefficients were as

follows: (a) Omega Within = 0.677 [C.I. = 0.648-0.707], Omega Between = 0.969 [C.I. = 0.950-0.989], (b) Maximal reliability *H* Within = 0.731 [C.I. = 0.709-0.755], *H* Between = 0.970 [C.I. = 0.951-0.990] (c) α_{pc} Within = 0.527, α_{pc} Between = 0.744, (d) Cronbach's alpha Within = 0.673 [C.I. = 0.631-0.718], Cronbach's alpha Between = 0.981 [C.I. = 0.431-1.00], (e) Cronbach's alpha with Kristof's correction, Within = 0.674 [C.I. = 0.632-0.719], Between = 0.981 [C.I. = 0.431-1.00], (f) lambda 1 Within = 0.505 [C.I. = 0.475-0.537], Between = 0.735 [C.I. = 0.322-1.00], and (g) lambda 2 Within = 0.675 [C.I. = 0.633-0.720], Between = 0.987 [C.I. = 0.432-1.00]. The obvious conclusion was that the estimates of the within level (person) were similar to the aggregate estimates but the point estimates of the between level were much higher, but with much less precision, likely because of the relatively small number of the level-2 units (universities in the present case). Figure 3 displays the distribution of within and between level coefficients following 1,000 replications using estimates of means and variances from the original dataset.

Conclusion

The purpose this study was to evaluate within and between level internal consistency estimates for a General Teacher test using various indices such as alpha, omega, maximal *H*, lambda 1, lambda 2, α_{pc} and a few corrections on them as past research has indicated that ignoring nesting can be detrimental to both parameter estimation, standard error estimation and consequently, reliability (Pornprasertmanit et al., 2014). The paper attempted to involve a wide variety of internal consistency indices including the early lower bound indices (Cronbach's alpha and its variants). Several important findings emerged in relation to measuring internal consistency reliability in multilevel versus aggregate structures.

The most important finding related to the measurement of reliability in that, differences in reliability at each level of the analysis suggests different levels of precision of the measured instrument. The present study included alpha reliability, composite reliability, maximal reliability, the α_{pc} statistic, and two of Guttman's popular lambda indices, namely λ_1 and λ_2. All suggested that the measurement of general competencies was more accurate and consistent at the university level (between university level of analysis Level-2) compared to the between-person level of analysis (Level-1). These findings suggest that, again, the aggregate measurement of reliability when ignoring nested structures can lead to misleading estimates regarding a measure's internal consistency estimate as the aggregate terms average the true reliabilities at each level. This would hinder the true

reliability of a measure with unknown consequences such as concluding unreliability, as would be the case in the present study for which between person estimates were low and, at times, unacceptable compared to the estimates derived at the university-level in the analysis for which consistency was remarkably high. It is important to note here, however, that different internal consistency estimates pose different assumptions regarding the measure under study and, thus, it will be important to evaluate the measure first and then select the most appropriate reliability estimate for the measure. For example, alpha assumes tau equivalence, an assumption that likely did not hold with the present data (see estimates of factor loadings in Figure 1).

The disparate findings regarding estimates of internal consistency reliability at the different levels in the analyses also question the earlier recommendations that ignored the higher order level is detrimental only under conditions of large ICCs (Kim, Kwok, & Yoon, 2012; Moerbeek, 2004; Pornprasertmanit et al., 2014). We found the opposite in the present study, in that small but non-negligible ICCs were associated with remarkably different coefficients at the different levels in the analyses. Thus, this earlier recommendation has been challenged with the present findings.

Differences between coefficients were also apparent. Alpha and its corrections, as well as Guttman's lambda coefficients, performed similarly and as lower bound estimates were also on the low side at the within person analysis, suggesting imprecise measurement at the person level. Neither point estimates nor their confidence intervals exceeded a recommended cutoff value of 0.80. In the presence of tau equivalence, as was the measure in the present study, omega and *H* were the most appropriate indices (Novick & Lewis, 1967), and they clearly suggested a better precision at the university level compared to the person level. Thus, scores across math departments tend to be more homogeneous compared to scores within departments. Further, structural models need to account for that level of precision when including covariates at the within level, which in the present study suggest that they are appropriate only at the university level.

The findings are limited for several reasons. First, the number of level-2 units was relatively small compared to what has been recommended in Monte Carlo simulation studies (Meuleman & Billiet, 2009). As Reise, Ventura, Nuechterlein, and Kim (2005) suggested, with few clusters the interpretation of factors can be difficult in light of the estimation involving aggregate terms. Meuleman and Billiet (2009) suggested that the number of clusters should be at a minimum 40 and approximately 60 if large structural effects are to be detected. However, due to the presence of a large number of units within clusters ($n = 87$), it has been

recommended that these large numbers compensate, to an extent, for the limited number of clusters as they found to be associated with smaller standard errors (Cohen, 1998; Hox & Maas, 2001; Snijders & Bosker, 1993). Second, our methodology for computing confidence intervals is only one among several possibilities; Padilla and Divers (2013) presented 6 different methodologies. Third, the estimates of internal consistency reliability used in the present study represent only a small fraction of the available estimates (Schmidt, Le, & Ilies, 2003). One motivation towards including some of these indices was ease to model them in Mplus compared to more cumbersome indices, such as split half estimates for which the number of possible splits with large numbers of items increases exponentially. Last, in the present study we ignored the influence of correlated errors, which may be detrimental for some coefficients compared to others; e.g., for alpha, which results in inflation (Komaroff, 1997).

It is important to assess how the above reliability coefficients behave when the data are dichotomous or polytomous (Dimitrov, 2002, 2003a, 2003b; Padilla & Divers, 2013) and not continuous as in the present study; see also Yang and Green (2014). The need to include modeling at various levels when the data are clustered is nevertheless imperative in light of the recent findings which show that ignoring clustering is associated with high Type-I error rates when assessing non-invariance (Kim et al., 2012) or the underestimation of standard errors (and, thus, Type-I error inflation) when covariates are modeled at the between level (Finch & French, 2011). Furthermore, it will be important to evaluate reliability in light of the properties of the measure (e.g., congeneric, tau equivalent, etc.) with the goal of selecting the most appropriate estimate for the data given evidence that reliability is often misconducted (Aiken et al., 1990). Last, corrective actions may need to be taken so that measurement error would be accounted for at each level in the analysis, prior to moving to more complex structural models using either Bayes' priors or information from past research; see G. Woodhouse et al. (1996) for a correction for unreliability. Other recommended approaches involve parcels to correct for unreliability (Coffman & MacCallum, 2005).

References

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., Roediger III, H. L., Scarr, S.,…& Sherman, S. J. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist, 45*(6), 721-734. doi: 10.1037/0003-066x.45.6.721

Anderson, J. C., & Gerbing, D. W. (1982). Some methods for respecifying measurement models to obtain unidimensional construct measurement. *Journal of Marketing Research, 19*(4), 453-460. doi: 10.1037/0003-066x.45.6.721

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) (pp. 1-19). New York, NY: Elsevier. doi: 10.1016/s1871-0301(06)01001-8

Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences, 12*(3), 291-294. doi: 10.1016/0191-8869(91)90115-r

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.

Brown, T. (2015). *Confirmatory factor analysis for applied research*. New York: Guilford.

Coffman, D. L., & MacCallum, R. C. (2005). Using parcels to convert path analysis models into latent variable models. *Multivariate Behavioral Research, 40*(), 235-259. doi: 10.1207/s15327906mbr4002_4

Cohen, M. (1998). Determining sample sizes for surveys with data analyzed by hierarchical linear models. *Journal of Official Statistics, 14*(3), 267-275.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104. doi: 10.1037/0021-9010.78.1.98

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. doi: 10.1007/bf02310555

Cronbach, L. J., & Gleser, G. C. (1964). The signal/noise ratio in the comparison of reliability coefficients. *Educational and Psychological Measurement, 24*(3), 467-480. doi: 10.1177/001316446402400303

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement, 62*(5), 783-801. doi: 10.1177/001316402236878

Dimitrov, D. M. (2003a). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement, 27*(6), 440-458. doi: 10.1177/0146621603258786

Dimitrov, D. M. (2003b). Reliability and true-score measures of binary items as a function of their Rasch difficulty parameter. *Journal of Applied Measurement, 4*(3), 222-233.

Dunn, T. J., Baguley, T., & Brunsden, V. (2013). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399-412. doi: 10.1111/bjop.12046

Feldt, L. S. (2002). Estimating the internal consistency reliability of tests composed of testlets varying in length. *Applied Measurement in Education, 15*(1), 33-48. doi: 10.1207/s15324818ame1501_03

Fife, D. A., Mendoza, J. L., & Terry, R. (2012). The assessment of reliability under range restriction. A comparison of α, ω, and test-retest reliability for dichotomous data. *Educational and Psychological Measurement, 72*(5), 862-888. doi: 10.1177/0013164411430225

Finch, W. H., & French, B. F. (2011). Estimation of MIMIC model parameters with multilevel data. *Structural Equation Modeling, 18*(2), 229-252. doi: 10.1080/10705511.2011.557338

Geldof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72-91. doi: 10.1037/a0032138

Goldstein, H. (2003). *Multilevel Statistical Models* (3rd ed.). London, UK: Arnold.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66*(6), 930-944. doi: 10.1177/0013164406288165

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255-282. doi: 10.1007/bf02288892

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.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164. doi: 10.1177/014662168500900204

Hayashi, K., & Kamata, A. (2005). A note on the estimator of the alpha coefficient for standardized variables under normality. *Psychometrika, 70*(3), 579-586. doi: 10.1007/s11336-001-0888-1

Heck, R. H. (1999). Multilevel modeling with SEM. In S. L. Thomas & R.H. Heck (Eds.), *Introduction to multilevel modeling techniques* (pp. 89-127). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques*. New York, NY: Routledge.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*(3), 177-189.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60*(4), 523-531. doi: 10.1177/00131640021970691

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling, 8*(2), 157-174. doi: 10.1207/s15328007sem0802_1

Huang, P.-H., & Weng, L.-J. (2012). Estimating the reliability of aggregated and within-person centered scores in ecological momentary assessment. *Multivariate Behavioral Research, 47*(3), 421-441. doi: 10.1080/00273171.2012.673924

Jackson, P., & Agunwamba, 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*(4), 567-578. doi: 10.1007/BF02295979

Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). *Research methods in social relations* (6th ed.). New York, NY: Harcourt Brace Jovanovich.

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling, 8*(3), 325-352. doi: 10.1207/s15328007sem0803_1

Kim, E. S., Kwok, O., & Yoon, M. (2012). Testing factorial invariance in multilevel data: A Monte Carlo study. *Structural Equation Modeling, 19*(2), 250-267. doi: 10.1080/10705511.2012.659623

Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.

Komaroff, E. (1997). Effect of simultaneous violations of essential τ-equivalence and uncorrelated error on coefficient *α*. *Applied Psychological Measurement, 21*(4), 337-348. doi: 10.1177/01466216970214004

Kopalle, P. K. (1997). Alpha inflation? The impact of eliminating scale items on Cronbach's alpha. *Organizational Behavior and Human Decision Processes, 70*(3), 189-197. doi: 10.1006/obhd.1997.2702

Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika, 28*(3), 221-238. doi: 10.1007/bf02289571

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(2), 395-407. Retrieved from http://www.jstor.org/stable/2345969

Li, H. (1997). A unifying expression for the maximal reliability of a linear composite. *Psychometrika, 62*(2), 245-249. doi: 10.1007/bf02295278

Liu, Y., Wu, A., & Zumbo, B. (2010). The impact of outliers on Cronbach's coefficient alpha estimate of reliability: Ordinal/rating scale item responses. *Educational and Psychological Measurement, 70*(1), 5-21. doi: 10.1177/0013164409344548

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86-92. doi: 10.1027/1614-2241.1.3.86

Marsh, H. W. (2007). *Self-concept theory, measurement and research into practice: The role of self-concept in educational psychology* (25th Vernon-Wall lecture series). London, UK: British Psychological Society.

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology, 23*(1), 1-21. doi: 10.1111/j.2044-8317.1970.tb00432.x

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum. doi: 10.4324/9781410601087

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods, 3*(1), 45-58. Retrieved from https://ojs.ub.uni-konstanz.de/srm/article/view/666

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*(1), 129-149. doi: 10.1207/s15327906mbr3901_5

Morin, A. J. S., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *Journal of Experimental Education, 82*(2), 143-167. doi: 10.1080/00220973.2013.769412

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*(4), 557-585. doi: 10.1007/bf02296397

Muthén, B. O. (1990). *Mean and covariance structure analysis of hierarchical data* (UCLA statistics series, #62). Los Angeles, CA: University of California, Los Angeles.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*(4), 338-354. doi: 10.1111/j.1745-3984.1991.tb00363.x

Muthén, B. O. (1994).Multilevel covariance structure analysis. *Sociological Methods & Research, 22*(3), 376-398. doi: 10.1177/0049124194022003006

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 267-316). Oxford, England: Blackwell. doi: 10.2307/271070

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*(1), 1-13. doi: 10.1007/bf02289400

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Opdenakker, M.-C., & Van Damme, J. (2000). Effects of schools, teaching staff and classes on achievement and well-being in secondary education: Similarities and differences between school outcomes. *School Effectiveness and School Improvement, 11*(2), 165-196. doi: 10.1076/0924-3453(200006)11:2;1-q;ft165

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5*(3), 343-355. doi: 10.1037/1082-989x.5.3.343

Padilla, M. A., & Divers, J. (2013). Bootstrap interval estimation of reliability via coefficient omega. *Journal of Modern Applied Statistical Methods, 12*(1), 78-89. doi: 10.22237/jmasm/1367381520

Pornprasertmanit, S., Lee, J., & Preacher, K. J. (2014). Ignoring clustering in confirmatory factor analysis: Some consequences for model fit and standardized parameter estimates. *Multivariate Behavioral Research, 49*(6), 518-543. doi: 10.1080/00273171.2014.933762

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures, 6*(2), 77-98. doi: 10.1080/19312458.2012.679848

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*(2), 173-184. doi: 10.1177/01466216970212006

Raykov, T. (1998). A method for obtaining standard errors and confidence intervals of composite reliability for congeneric items. *Applied Psychological Measurement, 22*(4), 369-374. doi: 10.1177/014662169802200406

Raykov, T. (2001). Bias of coefficient α for fixed congeneric measures with correlated errors. *Applied Psychological Measurement, 25*(1), 69-76. doi: 10.1177/01466216010251005

Raykov, T. (2002). Analytic estimation of standard error and confidence interval for scale reliability. *Multivariate Behavioral Research, 37*(1), 89-103. doi: 10.1207/s15327906mbr3701_04

Raykov, T. (2004). Estimation of maximal reliability: A note on a covariance structure modeling approach. *British Journal of Mathematical and Statistical Psychology*, *57*(1), 21-27. doi: 10.1348/000711004849295

Raykov, T. (2006). Interval estimation of optimal scores from multiple-component measuring instruments via SEM. *Structural Equation Modeling, 13*(2), 252-263. doi: 10.1207/s15328007sem1302_5

Raykov, T. (2012). Scale construction and development using structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 472-492). New York, NY: Guilford Press.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Routledge.

Raykov, T., & Marcoulides, G. A. (2012). Evaluation of validity and reliability for hierarchical scales using latent variable modeling. *Structural Equation Modeling, 19*(3), 495-508. doi: 10.1080/10705511.2012.687675

Raykov, T. Marcoulides, G., & Akaeze H. O. (2016). Comparing between- and within-group variances in a two-level study: A latent variable modeling approach to evaluating their relationship. *Educational and Psychological Measurement, 77*(2), 351-361. doi: 10.1177/0013164416634166

Raykov, T., & Penev, S. (2010). Evaluation of reliability coefficients for two-level models via latent variable analysis. *Structural Equation Modeling, 17*(4), 629-641. doi 10.1080/10705511.2010.510052

Raykov, T., & Pohl, S. (2012). On studying common factor variance in multiple-component measuring instruments. *Educational and Psychological Measurement, 73*(2), 191-209. doi: 10.1177/0013164412458673

Raykov, T., Rodenberg, C., & Narayanan, A. (2015). Optimal shortening of multiple-component measuring instruments: A latent variable modeling

procedure. *Structural Equation Modeling, 22*(2), 227-235. doi: 10.1080/10705511.2014.935927

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling, 9*(2), 195-212. doi: 10.1207/s15328007sem0902_3

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment, 84*(2), 126-136. doi: 10.1207/s15327752jpa8402_02

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(3), 351-357. doi: 10.2307/2087176

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8*(2), 206-224. doi: 10.1037/1082-989x.8.2.206

Schwartz, S. (1994). The fallacy of the ecological fallacy: The potential misuse of a concept and the consequences. *American Journal of Public Health, 84*(5), 819-824. doi: 10.2105/ajph.84.5.819

Shevlin, M., Miles, J. N. V., Davies, M. N. O., & Walker, S. (2000). Coefficient alpha: A useful indicator of reliability? *Personality and Individual Differences, 28*(2), 229-237. doi: 10.1016/s0191-8869(99)00093-8

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18*(3), 237-259. doi: 10.3102/10769986018003237

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72-101. doi: 10.2307/1412159

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271-295. doi: 10.1111/j.2044-8295.1910.tb00206.x

Streiner, D. L. (2003a). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80*(3), 217-222. doi: 10.1207/s15327752jpa8003_01

Streiner, D. L. (2003b). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99-103. doi: 10.1207/s15327752jpa8001_18

ten Berge, J. M. F., & Hofstee, K. B. (1999). Coefficients alpha and reliabilities of unrotated and rotated components. *Psychometrika, 64*(1), 83-90. doi: 10.1007/BF02294321

ten Berge, J. M. F., & Zegers, F. E. (1974). A series of lower bounds to the reliability of a test. *Psychometrika, 43*(4), 575-579. doi: 10.1007/bf02293815

Thompson, B., Green, S., & Yang, Y. (2010). Assessment of the maximal split-half coefficient t estimate reliability. *Educational and Psychological Measurement, 70*(2), 232-251. doi: 10.1177/0013164409355688

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. Chichester, UK: Wiley. doi: 10.1002/9781118356258

Werts, C. E., Linn, R. L., & Jöreskog, K. G. (1974). Intraclass reliability estimates: Testing structural assumptions. *Educational and Psychological Measurement, 34*(1), 25-33. doi: 10.1177/001316447403400104

Westfall, P. H., Henning, K. S. S., & Howell, R. D. (2012). The effect of error correlation on interfactor correlation in psychometric measurement. *Structural Equation Modeling, 19*(1), 99-117. doi: 10.1080/10705511.2012.634726

Wilhelm, P., & Schoebi, D. (2007). Assessing mood in daily life: Structural validity, sensitivity to change, and reliability of a short-scale to measure three basic dimensions of mood. *European Journal of Psychological Assessment, 23*(4), 258-267. doi: 10.1027/1015-5759.23.4.258

Woodhouse, G., Yang, M., Goldstein, H., & Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 159*(2), 201-212. doi: 10.2307/2983168

Woodhouse, K. (2015, June 12). Impact of Pell surge: Federal spending has overtaken state spending as the main source of public funding in higher education. *Inside Higher Ed*. Retrieved from: https://www.insidehighered.com/news/2015/06/12/study-us-higher-education-receives-more-federal-state-governments

Yang, Y., & Green, S. (2015). Evaluation of structural equation modeling estimates of reliability for scales with ordered categorical items. *Methodology, 11*(1), 23-34. doi: 10.1027/1614-2241/a000087

Yuan, K.-H., & Bentler, P. M. (2007). Multilevel covariance structure analysis by fitting multiple single-level models. *Sociological Methodology, 37*(1), 53-82. doi: 10.1111/j.1467-9531.2007.00182.x

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*(1), 123-133. doi: 10.1007/s11336-003-0974-7