METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

A Meta-Analysis of Construct Reliability Indices and Measurement Model Fit Metrics

Robert A. Peterson[a], Yeolib Kim[b], Boreum Choi[c]

[a] *Department of Marketing, The University of Texas at Austin, Austin, TX, USA.* **[b]** *School of Business Administration, Ulsan National Institute of Science and Technology, Ulsan, South Korea.* **[c]** *School of Business Administration, University of Seoul, Seoul, South Korea.*

Methodology, 2020, Vol. 16(3), 208–223, https://doi.org/10.5964/meth.2797

Received: 2019-03-10 • **Accepted:** 2019-10-16 • **Published (VoR):** 2020-09-30

Corresponding Author: Yeolib Kim, School of Business Administration, Ulsan National Institute of Science and Technology, 50 UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan, South Korea 44919. Tel. +82 10 3533 4986, E-mail: yeolib.kim@unist.ac.kr

Abstract

The present research examined the distributional properties of construct reliability indices and model fit metrics, explored relationships between and among the indices and metrics, and investigated variables influencing the relative magnitudes of the indices and metrics in structural equation measurement models. A broad-based meta-analysis of reported construct reliability indices and selected model fit metrics revealed modest relationships among reliability indices, minimal relationships among model fit metrics, and a virtual absence of relationships between reliability indices and model fit metrics. Differences in magnitudes of selected reliability indices and model fit metrics were found to primarily be a function of the (total) number of items employed in a measurement model. The implications of the findings suggest that the current practice of indiscriminately computing and reporting of reliability indices and model fit metrics based only on arbitrary heuristics should be abolished and replaced by theoretically justified indices and metrics.

Keywords

structural equation model, measurement model, confirmatory factor analysis, construct reliability, model fit, meta-analysis

In a seminal article, Anderson and Gerbing (1988) recommended that evaluating structural equation models should consist of two distinct steps. The first step should be an evaluation of the underlying measurement model (or confirmatory factor analysis), whereas the second should be an evaluation of the structural model. As a consequence of their article, virtually all structural equation models applied by behavioral researchers have adopted this recommended approach.

Perusal of the empirical literature reveals that most evaluations of measurement models now begin by assessing the reliability of the constructs being investigated in conjunction with the evaluation of the measurement model itself. If the properties of measures mapping into constructs or latent variables are deemed to possess "adequate" reliability, usually by comparing calculated reliability indices to some heuristically determined criteria, the measurement model itself is then assessed to determine whether it is "acceptable" (Menold, Bluemke, & Hubley, 2018). Analogous to the assessment of reliability, whether a measurement model is "acceptable" is determined by calculating one or more model fit metrics and comparing them to heuristic criteria (e.g., Bagozzi & Yi, 1988).

It is generally asserted that the reliability of a measurement model's constructs and the fit of the model itself are distinct and therefore have to be separately satisfied to validate the measurement model (e.g., Bagozzi & Yi, 1988). However, there is a dearth of studies addressing this assertion of independence, and the studies that support the assertion (e.g., Stanley & Edwards, 2016) tend to be simulations or use artificial data.

Given that several different reliability indices and model fit metrics are typically calculated and reported when evaluating a measurement model, the purpose of the present research was to address the question, "Is there an empirical relationship between construct reliability indices and model fit metrics in measurement models?" Further, when addressing this question, the present research also addressed two prefatory questions: "Are there empirical relationships among common construct reliability indices in a measurement model?" and "Are there empirical relationships among common model fit metrics in a measurement model?" A meta-analytic approach based on (non-simulated) data harvested from reported measurement models was employed to answer the three questions. Based on the answers, suggestions as to how to improve the application of measurement models are discussed. As such, findings from the research have broad implications for evaluating, interpreting, and reporting measurement models (e.g., Menold et al., 2018).

Relationships Between Reliabilities and Model Fit Metrics

Three indices are commonly used to measure the reliability of constructs or latent variables in structural equation measurement models: coefficient alpha, composite reliability (CR), and average variance extracted (AVE). Commonly recommended acceptance thresholds are .70 for alpha and CR, and .50 for AVE (Fornell & Larcker, 1981). In general, alpha is the most widely used reliability index, and even though it has been criticized as possessing numerous deficiencies (e.g., Cho & Kim, 2015; McNeish, 2018; Schmitt, 1996), estimates of "true reliability" based on composite reliability are, on average, similar to those based on coefficient alpha (Peterson & Kim, 2013). The magnitudes of alpha, CR, and AVE are related in that if one is "large," then the others will also tend to be "large."

However, "large" alpha and CR values do not always guarantee "large" AVE values. For example, Ping (2004) demonstrated that while reducing the number of items for a construct can increase coefficient alpha and composite reliability values, doing so can simultaneously reduce average variance extracted values. Relative differences between the magnitudes of alpha, CR, and AVE values depend on a variety of factors, including research design characteristics and the quality of measured items (Ping, 2004).

Evaluating the fit of a structural equation measurement model involves a statistical comparison of the model-implied population covariance and the observed covariance adjusting for sample size, number of constructs, and/or degrees of freedom (Hooper, Coughlan, & Mullen, 2008). The present research focuses on four commonly applied absolute fit metrics, the relative/normed χ^2 statistic, χ^2/df; the root mean square error of approximation (RMSEA); the goodness-of-fit index (GFI); and the standardized root mean square residual (SRMR); and one incremental fit metric, the comparative fit index (CFI). These metrics are representative of model fit metrics (Hu & Bentler, 1999; West, Taylor, & Wu, 2012). Recommended acceptance thresholds of .07, .90, .08, and .95 constitute the general consensus for, respectively, RMSEA, GFI, SRMR, and CFI (Hooper et al., 2008). Recommended acceptance thresholds for χ^2/df range from 2.0 to 5.0 (Hooper et al., 2008). Other metrics sometimes used are closely related to these five, and including them in the analysis would be redundant.[1]

As recommended by Hooper et al. (2008), it is common practice to report multiple model fit metrics when evaluating a measurement model, primarily because each model fit has its unique strengths and flaws. For instance, RMSEA is useful for penalizing overfitting models but it underestimates model fit when N is smaller than 200 (Curran, Bollen, Chen, Paxton, & Kirby, 2003). Despite the plethora of literature covering the list of model fits metrics, it is still unclear how model fit metrics are related to each other in actual applications of measurement models. A likely explanation for the lack of understanding can be traced to how each model fit metric incorporates model weights (between the observed covariance and model variance), sample size, and model complexity (Barrett, 2007). Hence, to improve the application and interpretation of measurement models, it is important to identify "true" relationships among model fit metrics using non-artificial data to provide clarity and guidance on model fit evaluation and reporting.

The possible relationship or lack thereof between reliability indices and model fit metrics remains an intriguing and relatively under-discussed topic. To illustrate, whereas Stanley and Edwards (2016) concluded there was no relationship between the two based on their simulation results, Kanyongo and Schreiber (2009) found that "the smaller the alpha, the better the model fit" for factor analyses in their simulation study (p. 466).

1) Interested readers can contact the authors for the results of the other model fit metrics (e.g., AGFI, NFI). The other model fit metrics had a correlation of at least .75 with one of the investigated model fit metrics. Consequently, these model fit metrics were dropped from the analyses.

Reliabilities are based on factor loadings in a measurement model, and obtaining certain high/low reliabilities can signal that a model is specified correctly or incorrectly (e.g., items loading onto the wrong construct), thereby affecting model fit metrics (Fan & Sivo, 2007; Kanyongo & Schreiber, 2009). Using actual, non-simulated data to assess the relationship between reliability indices and model fit metrics is critical for understanding the "true" psychometric properties of measurement models.

Method

Articles reporting an empirical reflective structural equation modeling analysis application and containing one or more alpha, CR, AVE, and/or model fit metric values served as the data source for the present research. Terms such as "measurement model," "confirmatory factor analysis," "structural equation," "(coefficient) alpha," "composite reliability," "average variance extracted," "model fit," and a combination of these terms were searched in Business Source Premier, Communication & Mass Media Complete, JSTOR, PsycARTICLES, PsycINFO, ScienceDirect, and Springer Link databases.

Then, to provide a broad representation of measurement models, an issue-by-issue search was made using the same terms in several prominent journals in psychology, marketing, business, education, information systems, and management. The search covered the period 1996 through 2017. Five hundred fifty-seven articles were initially identified that contained potentially usable measurement model data. Studies applying a partial least squares (PLS) technique were excluded as it is not a covariance-based modeling technique. One hundred-fifty articles were eliminated due to this exclusion criterion.

The final data base consisted of 312 articles reporting 332 studies drawn from 93 journals. The total sample size was 243,700 individuals, with a mean of 734 individuals per study. On average, there were 6.39 constructs per study. All harvested reliability indices and model fit metrics were retained for analysis, even if they were possible outliers. Construct reliability indices were based solely on reflective constructs in a measurement model. On average, the number of reliability indices reported per construct was 2.1, and the number of fit metrics reported per model was 5.3. Based on the harvested data, measures of central tendency, measures of variability, and pairwise correlations respectively among and between reliability indices and model fit metrics were computed.

In addition to these measures, differences among reliability indices (Peterson & Kim, 2013) and differences among four model fit metrics (Rigdon, 1996) were computed to facilitate quantitative comparisons of the measures. (χ^2/df was excluded from the fit metric computations since it is not based on a zero-to-one scale.) Since a value of 1.0 would imply a perfect fit for CFI and GFI whereas a value of 0.0 would imply a perfect fit for RMSEA and SRMR, following Lai and Green (2016), values for RMSEA and SRMR were rescaled into (1 – RMSEA) and (1 – SRMR). Measures of central tendency and measures of variability were computed for (CR – alpha), (CR – AVE), (alpha –

AVE), (CFI – RMSEA), (CFI – GFI), (CFI – SRMR), (GFI – SRMR), (GFI – SRMR), and (RMSEA – SRMR). Respective differences served as dependent variables in a series of regression analyses with sample size and number of items linked to a construct serving as independent variables for the reliability indices, and sample size, (total) number of items in the measurement model, and (total) number of constructs in the measurement model serving as independent variables for the model fit metrics (Raykov, 1998).

Results

Distributional Properties

Table 1 provides the distributional properties of the reliability indices and model fit metrics examined. Note that the number of observations varied across the reliability indices and model fit metrics; they ranged from 1,552 for CR to 72 for SRMR. The mean alpha was .85; the mean CR was .86; and the mean AVE was .69. The mean CR was slightly larger than the mean alpha, confirming past research (Peterson & Kim, 2013) but implying criticisms of alpha may be overstated from a practical perspective. In general, the mean alpha and CR values were respectively larger than the mean AVE value by .16 (23% difference) and .17 units (25% difference). The alpha (sk = -1.02) and CR (sk = -0.80) distributions were somewhat negatively skewed, whereas the AVE (sk = 0.02) distribution was approximately symmetric. Moreover, the alpha (k = 1.74) and CR (k = 0.48) distributions were slightly leptokurtic, whereas the AVE distribution was slightly platykurtic (k = -0.60).

Table 1

Descriptive Statistics of Reliability Indices and Model Fit Metrics

Characteristic	α	CR	AVE	χ^2/df	CFI	RMSEA	GFI	SRMR
N	1.115	1.552	1.522	246	253	227	143	72
M	.85	.86	.69	4.01	.95	.06	.91	.05
Mdn	.86	.88	.69	2.01	.96	.06	.91	.05
SD	.08	.07	.12	17.23	.03	.02	.04	.02
Minimum	.38	.53	.31	.89	.77	.00	.73	.02
Maximum	.99	.99	.99	249.97	1.00	.13	.99	.15
Range	.61	.46	.68	249.08	.23	.13	.26	.13
Skewness	-1.02	-.80	-.02	12.73	-1.47	.10	-.63	2.07
Kurtosis	1.74	.48	-.60	174.66	4.76	1.67	1.31	9.26

Note. CR = Composite reliability; AVE = Average variance extracted; CFI = Comparative fit index; RMSEA = Root mean square error of approximation; GFI = Goodness-of-fit index; SRMR = Standardized root mean square residual.
$^{*}p < .05$. $^{**}p < .01$.

Majorities of the reported values for alpha and composite reliability were above the recommended (heuristic) thresholds. For alpha, 77% of the reported values exceeded .80, 96% exceeded .70, and 99% exceeded .60. For CR, 83% of the reported values exceeded .80, 98% exceeded .70, and 99% exceeded .60. Ninety-six percent of the reported AVE values were greater than the recommended threshold of .50. In studies where all three reliability index values were reported, approximately 93% of the alpha, composite reliability, and average variance extracted values were jointly above the heuristic criterion of .70 for alpha and CR and .50 for AVE. Hence, based on the heuristic criteria, the magnitudes of the construct reliability indices in reported structural equation measurement models were generally adequate. Two percent of the reliability indices consisted of alpha and CR values, but not AVE values, meeting the heuristic criteria.

With respect to the model fit metrics, the mean χ^2/df was 4.01; the mean CFI was .95; the mean RMSEA was .06; the mean GFI was .91; and the mean SRMR was .05. The χ^2/df (sk = 12.73) and SRMR (sk = 2.07) distributions were significantly positively skewed; the CFI (sk = -1.47) and GFI (sk = -0.63) distributions were somewhat negatively skewed; and the RMSEA (sk = 0.10) distribution was approximately symmetrical. The χ^2/df (k = 174.66), CFI (k = 4.76), and SRMR (k = 9.26) distributions were highly leptokurtic; the RMSEA (k = 1.67) and GFI (k = 1.31) distributions were somewhat leptokurtic. Given these distributional properties, caution should be exercised when interpreting model fit metrics such as CFI and SRMR due to the existence of significant outliers.

A majority of the reported model fit metric values met the recommended (heuristic) thresholds. For χ^2/df, 49% were smaller than 2.00, 83% were smaller than 3.00, and 96% were smaller than 5.00. Sixty-eight percent of the CFI values were larger than .95. For RMSEA, 51% were smaller than .06, 71% were smaller than .07, and 89% were smaller than .08. Sixty-eight percent of the GFI values exceeded the GFI threshold criterion of .90. Ninety-two percent of the SRMR values were less than .08. In general, reported measurement model fit metric values were acceptable according to traditional heuristic criteria. χ^2/df and GFI have come under scrutiny for their sensitivity to sample size (Barrett, 2007). When χ^2/df and GFI were excluded, 69% of the model fit metric values collectively exceeded the respective thresholds for studies that reported CFI, RMSEA, and SRMR.

Index and Metric Relationships

Three sets of pairwise correlations were computed based on reported construct reliability indices and model fit metrics: (i) within-set correlations between comparable reliability indices, (ii) within-set correlations between comparable model fit metrics, and (iii) cross-set correlations between comparable reliability indices and model fit metrics. Table 2 contains these correlations. The correlations for alpha and CR, alpha and AVE, and CR and AVE were respectively .78, .62, and .71. Hence, the reliability pairs exhibited approximately 61%, 38%, or 50% in shared variation.

Table 2

Correlations of Reliability Indices and Model Fit Metrics

Metric	1	2	3	4	5	6	7
1. α	-						
2. CR	.78** (697)	-					
3. AVE	.62** (786)	.71** (1.388)	-				
4. χ^2/df	.02 (785)	.09** (1.242)	.06 (1.208)	-			
5. CFI	.08* (816)	.05 (1.272)	.13** (1.245)	-.08 (232)	-		
6. RMSEA	-.02 (782)	.00 (1.147)	.01 (1.130)	.23** (211)	-.43** (218)	-	
7. GFI	-.09* (471)	-.14** (748)	-.06 (707)	.06 (132)	.50** (129)	-.36** (123)	-
8. SRMR	-.09 (241)	-.11* (353)	-.11* (373)	-.06 (64)	-.48** (69)	.22 (64)	-.17 (27)

Note. Values in parenthesis indicate number of matched samples. CR = Composite reliability; AVE = Average variance extracted; CFI = Comparative fit index; RMSEA = Root mean square error of approximation; GFI = Goodness-of-fit index; SRMR = Standardized root mean square residual.
*$p < .05$. **$p < .01$.

Although five of the 10 correlations among the model fit metrics were statistically significant at the .05 significance level, they were relatively small in magnitude. In particular, they ranged from |.06| to |.50|, with a median value of |.23|. Thus, the model fit metrics had pairwise shared variances ranging from zero to 25%, with a median of 5%. Further, inspection of Table 2 reveals that there was no pattern to the directionality of the model fit metric correlations, with certain model fit metrics correlating both positively and negatively with other model fit metrics.[2]

The 15 pairwise correlations between construct reliability indices and model fit metrics were minimal to nonexistent. As can be seen in Table 2, only four reached double-digits. The reliability index-model fit metric correlations ranged from |.00| (CR & RMSEA) to a maximum of |.14| (CR & GFI), with the latter representing a shared variance of 2%. The median correlation was |.06|.

In those instances wherein reported alpha, CR, and AVE values all exceeded the recommended reliability threshold criteria, 4.5%, 41.3%, 10.6%, 41.2%, and 1.4% of the studies respectively reporting χ^2/df, CFI, RMSEA, GFI, and SRMR values did not meet the most conservative recommended model fit metric acceptance threshold criteria. There were five studies in which none of the reported model fit metrics satisfied the recommended threshold values. Among studies where reported χ^2/df, CFI, RMSEA, GFI, and SRMR values all exceeded the recommended model fit threshold criteria, 3.9% of reported alpha values and 2.3% of reported CR values did not meet the threshold criterion of .70, and

2) In Table 2, for every model fit value, there are multiple constructs. The *N*'s for correlations between reliability indices and model fit metrics are computed based on the *N*'s of the model fit metrics.

3.7% of reported AVE values did not meet the threshold criterion of .50. One study did not satisfy any of the recommended reliability threshold criteria.

Because there were minimal or nonexistent relationships among the measurement model fit metrics, regression analyses were first carried out to determine whether there was a relationship between the magnitude of a fit metric and the size of the sample used, the total number of items linked to the constructs, and the number of constructs in a measurement model. Only the number of items had a significant, but not substantial, effect on χ^2/df (β = .51, R^2 = .03), CFI (β = -.01, R^2 = .04), GFI (β = -.01, R^2 = .04) or SRMR (β = .00, R^2 = .02).

Exploring Index and Metric Differences

Similarly, because there were at best moderate relationships between the construct reliability indices and minimal or nonexistent relationships between the measurement model fit metrics, analyses were undertaken to compare the respective magnitudes of the reliability indices and model fit metrics. Since the heuristic for χ^2/df was greater than an absolute 1.0, it was not included in this set of analyses. Specifically, within-sample t-tests were first conducted to determine whether mean reliability index and model fit metric differences were statistically significant. Then, regression analyses were undertaken to explore whether sample size and number of items in each construct related to differences between pairs of reliability indices, and whether sample size, (total) number of items in the measurement model, and (total) number of constructs in the measurement model related to differences between pairs of model fit metrics.

Table 3 contains descriptive statistics for the nine pairs of differences analyzed. With the exception of (CFI – SRMR), statistically significant differences were observed for all pairs of construct reliability indices and model fit metrics. However, apart from the (CR – AVE) and (alpha – AVE) comparisons, the differences, while significant, were not substantial.

Table 4 reports the estimated relationships between sample size, number of items, number of constructs, and the nine reliability index and fit metric differences examined. All variables were standardized before conducting the analyses. In general, Table 4 reveals that (i) sample size did not account for significant variability in the reliability index and fit metric differences; and (ii) only in two instances, (CFI – GFI) and (GFI – RMSEA), did the number of constructs account for moderate variation in the differences observed for the fit metrics. However, the total number of items contained in a measurement model was significantly related to differences for seven of the nine comparisons, and in certain instances (e.g., GFI – SRMR, GFI – RMSEA, and CR – AVE), the variance accounted for in the differences was substantial.

Table 3

Descriptive Statistics of Differences Among Reliability Indices and Model Fit Metrics

Characteristic	CR – α	CR – AVE	α – AVE	CFI – RMSEA	CFI – GFI	CFI – SRMR	GFI – RMSEA	GFI – SRMR	RMSEA – SRMR
N	697	1.388	786	218	129	69	123	27	64
M	.01**	.18**	.18**	.02**	.05**	.00	-.04**	-.05**	.01*
Mdn	.00	.18	.18	.02	.05	.01	-.04	-.04	.01
SD	.05	.09	.09	.03	.04	.03	.04	.05	.03
Minimum	-.20	-.07	-.13	-.17	-.07	-.11	-.23	-.14	-.08
Maximum	.29	.53	.50	.10	.18	.09	.05	.05	.09
Range	.49	.60	.63	.27	.25	.20	.28	.19	.17
Skewness	.57	.20	-.19	-1.54	.59	-.77	-.95	.06	.02
Kurtosis	5.92	.06	.23	7.99	2.21	3.82	3.31	-.13	3.02

Note. CR = Composite reliability; AVE = Average variance extracted; CFI = Comparative fit index; RMSEA = Root mean square error of approximation; GFI = Goodness-of-fit index; SRMR = Standardized root mean square residual.
*$p < .05$. **$p < .01$.

Table 4

Regression Analysis of Methodological Characteristics and Reliability Indices and Model Fit Metrics

	Sample size		Number of items		Number of constructs	
Pair	β	R^2 (%)	β	R^2 (%)	β	R^2 (%)
CR – α	.10**	1.0	-.02	0.0	-.17**	2.7
CR – AVE	.02	0.0	.47**	22.5	-.18**	3.4
α – AVE	-.07*	0.6	.39**	15.3	-.10**	0.9
CFI – RMSEA	.04	0.2	-.17**	4.1	-.08	0.7
CFI – GFI	-.13	1.8	.37**	14.8	.30**	9.0
CFI – SRMR	-.16	2.5	-.26*	6.7	.08	0.6
GFI – RMSEA	.10	1.0	-.43**	21.8	-.32**	10.3
GFI – SRMR	.14	2.1	-.51**	25.6	-.23	5.3
RMSEA – SRMR	.31*	9.7	-.08	0.8	-.18	3.3

Note. Number of items for reliability indices corresponds to the total number of items for each construct. Number of items for model fit metrics corresponds to the total number of items included in the measurement model. Number of constructs for model fit metrics corresponds to the total number of constructs included in the measurement model. CR = Composite reliability; AVE = Average variance extracted; CFI = Comparative fit index; RMSEA = Root mean square error of approximation; GFI = Goodness-of-fit index; SRMR = Standardized root mean square residual.
*$p < .05$. **$p < .01$.

Discussion and Conclusion

The present research systematically examined and analyzed actual construct reliability and model fit metric data from 332 studies reporting the results of applying measurement models. The majority of construct reliability indices and model fit metrics examined exceeded the respective heuristic thresholds that have been developed and refined across decades of construct measurement and structural equation modeling. This is to be expected given the typical reporting practices of researchers and the evaluation protocols of journal reviewers and editors when deciding whether a measurement model is acceptable or valid. In particular, based on traditional heuristic criteria, most measurement models reported in journal articles are assumed to be acceptable or valid in that construct reliabilities and model fits are both satisfied. If they were not acceptable, to either the researcher(s) submitting them for publication consideration or the reviewer(s) and editor(s) evaluating them, the measurement model would most likely not appear in a journal article.

Properly assessing and reporting reliability constructs and model fit metrics are imperative for evaluating, interpreting, and communicating the validity of a measurement model. Indeed, researchers are increasingly using construct reliability indices and model fit metrics as complementary measures when gauging the extent to which a measurement model is valid. However, until this study, there has been no documentation or comparison of the distributional properties of reported construct reliability indices alpha, CR, or AVE and model fit metrics χ^2/df, SRMR, RMSEA, GFI, or CFI, or their interrelationships. Thus one contribution of the present research was to provide a quantitative perspective on the distributional properties of reported construct reliability indices and model fit metrics as well as on the relationships respectively among and between construct reliability indices and model fit metrics.

Because the findings reported in this manuscript are based on a broad range of research domains and a diversity of measurement model applications, they possess generality. Consequently, the findings are amenable to being employed actuarially to both complement and supplement currently recommended threshold heuristics when assessing the "adequacy" of a measurement model to acquire a more nuanced evaluation of model adequacy. Such a use is consistent with the call for flexible threshold criteria when evaluating measurement models (e.g., Niemand & Mai, 2018).

The present research answered the initial three questions guiding the research in that it revealed that empirical relationships among the three construct reliability indices were modest at best, minimal among the five model fit metrics, and practically nonexistent between construct reliability indices and model fit metrics. In brief, the reliabilities of constructs in a measurement model and the associated model fit metrics examined were effectively independent and unrelated. As such, the research empirically corroborates conventional wisdom that construct reliability and measurement model fit are essentially distinct concepts.

The present research complements the research of Stanley and Edwards (2016), who have argued on the basis of artificial data that there should be no relationship between construct reliability indices and model fit metrics given that construct reliability is a property of a single construct whereas model fit is a property of a measurement model that contains multiple constructs. At the same time, though, the present research does not support the conclusion of Kanyongo and Schreiber (2009), also based on artificial data, that a negative relationship exists between construct reliability and model fit.

The present research likewise empirically demonstrated that while it is possible to obtain acceptable reliabilities but unacceptable model fit metrics, or acceptable model fit metrics but unacceptable reliabilities (e.g., Ping, 2004; Stanley & Edwards, 2016), in practice such situations appear to only rarely occur (see Table 2 for negative correlations between selected reliability indices and model fit metrics).

Another contribution of the present research is that it documented factors that contribute to differences in the magnitudes of construct reliability indices and model fit metrics, even those created to purportedly measure the same measurement model characteristic. Reliability indices and model fit metrics respectively quantify different aspects of reliability and model fit, and are in many instances based on different assumptions and have been created for different purposes (e.g., Lai & Green, 2016). Even so, researchers sometimes appear to indiscriminately calculate and report various reliability indices and model fit metrics without being aware of these distinctions.

In addition to demonstrating that the model fit metrics examined were measuring model fit from different perspectives, the present research provided evidence that fit metric values reported for a particular measurement model tended to be significantly different, even though similar cutoff criteria were frequently applied to the fit metrics. Importantly, the present research showed empirically that to a substantial extent differences in the magnitudes of model fit metric values were due to the total number of items incorporated in a measurement model. For example, the larger the number of items incorporated in a measurement model, the smaller the difference between the value of GFI relative to the value of SRMR, and the larger the value of CR relative to the value of AVE. Hence, for a comprehensive assessment of a measurement model, it is necessary to go beyond simply calculating and reporting goodness-of-fit measures and take into consideration research design characteristics. That the relative magnitudes of several of the indices and metrics investigated appeared to be sensitive to the (total) number of items incorporated in a measurement model requires replication and further investigation.

There is virtual consensus that measurement models must be based on theory and not merely be a consequence of "fishing," "dustbowl empiricism," HARKing, or "*p*-hacking" if they are to prove useful (e.g., Hooper et al., 2008). Hence, given the results of the present research, it is imperative that the indices and metrics used to evaluate a measurement model be consistent with the theory underlying the model, and reliability indices and

fit metrics not simply be reported and then ignored. Relatedly, based on the present research, it would seem that constructing or modifying a measurement model based only on construct reliabilities or model fit metrics without a strong theoretical foundation or rationale for doing so suggests poor research practice, especially if construct reliability indices and/or model fit metrics are calculated and then selectively reported (e.g., Straub, 1989).

Stated somewhat differently, assuming there is sound theory, although a full complement of reliability indices and model fit metrics are typically computed when constructing a measurement model, given the present results, only those that can be justified by the underlying theory should be assessed and reported. Indiscriminately reporting all measurement statistics provided by standard computer software programs (e.g., LISREL, AMOS, EQS, Mplus) without linking or justifying the use of individual indices or fit metrics to the theory underlying the measurement model is unwarranted.

Results from the construct reliability analysis revealed that alpha, CR, and AVE respectively had shared variances of 61%, 38%, and 50%. Hence there was unshared variance due to other causes (e.g., measurement error) as well as the fact that what they measure differs (e.g., Farrell, 2010). Thus, for example, depending on the theory underlying a construct, CR may be a more appropriate index if the emphasis is on the proportion of total variance explained, whereas AVE may be more appropriate if the emphasis is on common variance explained. Moreover, if discriminant validity is an issue, then AVE may be more appropriate than CR (Farrell, 2010).

Similarly, model fit metrics had pairwise shared variances ranging from none to 25%. Given the large unshared variances among the model fit metrics, different measurement model characteristics are being assessed by different fit metrics. Hence, discretion should be followed when selecting model fit metrics to compute, evaluate, and report on the appropriateness of a measurement model. Depending on the theory underlying the model and research characteristics such as sample size, number of items, data type, and model testing stage, certain model fit metrics are more appropriate than other model fit metrics (e.g., Barrett, 2007). If the research focus is on the discrepancy between a measurement model and a sample covariance matrix, some variant of chi-square is an appropriate metric. However, if the research focus is on the proportion of variance accounted for by an estimated population covariance matrix, then some variant of a goodness-of-fit metric is appropriate.

The rhetorical question raised by the present research relates to the broad issue of constructing and evaluating a measurement model. If there are relatively weak relationships between reliability indices and between model fit metrics, and virtually no relationship between the reliability of a measurement model (i.e., its constructs or latent variables) and the fit of the model, what and how should construct reliability indices and model fit metrics be assessed? What does it mean if the various reliability indices and model fit metrics are essentially unrelated? Stanley and Edwards (2016) discussed

situations in which construct reliability and model fit can be deemed "acceptable" or "unacceptable." If there is no relationship between construct reliability and measurement model fit, and the respective indices and metrics are internally inconsistent, then the notion of what constitutes a satisfactory measurement model needs to be revisited.

Limitations and Needed Research

This study is not without limitations. Analogous to any meta-analysis, contrary to the best intentions, it is possible that relevant studies were missed. And, despite the attempt to capture and quantitatively analyze measurement model data from original research articles, there is the possibility of publication bias that is inherent in quantitative reviews. This could be potentially significant given that, as mentioned previously, measurement models that are found to be "unacceptable" are not likely to appear in the literature. Such possible "truncation" may have influenced the relationships observed in the present research. Moreover, because the analyses were based on widely differing sample sizes, and in some instances rather small sample sizes, a more comprehensive multivariate analysis that could have produced more insights into the relationships between and among construct reliability indices and model fit metrics as well as between the indices and metrics and the research design characteristics investigated was not possible. Consequently, future research efforts should address these possible limitations.

Future research should also strive to better understand the relationship between the reliability of the constructs in a measurement model and the associated measurement model fit metrics. Too often it seems that reliability indices possess no decisional utility. They are simply calculated and reported as an afterthought with no assessment. For example, perusal of the literatures that produced the data for this study revealed an abundance of perfunctory reporting of reliability indices with no commentary on their assessment or application when evaluating the appropriateness of a measurement model.

Three possible extensions of the present study are (i) a further meta-analytic assessment of relationships between and among reliabilities and model fits that also includes factor loadings and selected research design characteristics other than sample size and number of items or constructs and takes into account additional disciplines; (ii) simulations incorporating different levels of measurement error and different test conditions (e.g., dropping items, re-specifying the measurement model) to compare their joint effects on construct reliability indices and model fit metrics; and/or (iii) one or more large-scale empirical studies specifically designed to take into account a variety of variables, constructs, measurement models, research design characteristics, and reliability indices and model fit metrics under controlled conditions. For instance, the impact that the number of items in a measurement model has on the relative magnitudes of the reliability indices and fit metrics requires further study to determine its origin or mechanism.

Moreover, analytical research should be undertaken to demonstrate and understand why construct reliability indices are not related to model fit metrics. The present re-

search and prior simulation research have documented the relationships and discussed the conceptual reasons for them. Analytical demonstrations would seem necessary to provide a transparent and comprehensive perspective to obtain closure on the question.

In conclusion, the present results reinforce the need for additional (global) statistics that take into account and simultaneously incorporate both construct reliability and model fit when assessing a measurement model (e.g., Heene, Hilbert, Draxler, Ziegler, & Buhner, 2011). Applying a measurement model wherein various reliability indices and model fit metrics are only weakly related internally and virtually independent cross-measures of reliability and model fit would seem to be a questionable research practice. Only by going beyond the use of singularly unique measures with (arbitrary) heuristic assessment thresholds will structural equation modeling fulfill its promise. Only through such endeavors will it be possible to fully understand the impact of construct reliabilities and model fit metrics on the quality of a measurement model.

Funding: The authors have no funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Acknowledgments: The authors have no support to report.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411-423. https://doi.org/10.1037/0033-2909.103.3.411

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 16*(1), 74-94. https://doi.org/10.1007/BF02723327

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42*(5), 815-824. https://doi.org/10.1016/j.paid.2006.09.018

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods, 18*(2), 207-230. https://doi.org/10.1177/1094428114555994

Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. B. (2003). Finite sampling properties of the point estimates and confidence intervals of the RMSEA. *Sociological Methods & Research, 32*(2), 208-252. https://doi.org/10.1177/0049124103256130

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42*(3), 509-529. https://doi.org/10.1080/00273170701382864

Farrell, A. M. (2010). Insufficient discriminant validity: A comment on Bove, Pervan, Beatty, and Shiu (2009). *Journal of Business Research, 63*(3), 324-327. https://doi.org/10.1016/j.jbusres.2009.05.003

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *JMR, Journal of Marketing Research, 18*(1), 39-50. https://doi.org/10.1177/002224378101800104

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Buhner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods, 16*(3), 319-336. https://doi.org/10.1037/a0024917

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53-60.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Kanyongo, G. Y., & Schreiber, J. B. (2009). Relationship between internal consistency and goodness of fit maximum likelihood factor analysis with varimax rotation. *Journal of Modern Applied Statistical Methods, 8*(2), 463-468. https://doi.org/10.22237/jmasm/1257034140

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research, 51*(2-3), 220-239. https://doi.org/10.1080/00273171.2015.1134306

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412-433. https://doi.org/10.1037/met0000144

Menold, N., Bluemke, M., & Hubley, A. M. (2018). Validity: Challenges in conception, methods, and interpretation in survey research. *Methodology, 14*(4), 143-145. https://doi.org/10.1027/1614-2241/a000159

Niemand, T., & Mai, R. (2018). Flexible cutoff values for fit indices in the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 46*(6), 1148-1172. https://doi.org/10.1007/s11747-018-0602-9

Peterson, R. A., & Kim, Y. (2013). On the relationship between coefficient alpha and composite reliability. *The Journal of Applied Psychology, 98*(1), 194-198. https://doi.org/10.1037/a0030767

Ping, R. A., Jr. (2004). On assuring valid measures for theoretical models using survey data. *Journal of Business Research, 57*(2), 125-141. https://doi.org/10.1016/S0148-2963(01)00297-1

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement, 22*(4), 375-385. https://doi.org/10.1177/014662169802200407

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling, 3*(4), 369-379. https://doi.org/10.1080/10705519609540052

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353. https://doi.org/10.1037/1040-3590.8.4.350

Stanley, L. M., & Edwards, M. C. (2016). Reliability and model fit. *Educational and Psychological Measurement, 76*(6), 976-985. https://doi.org/10.1177/0013164416638900

Straub, D. W. (1989). Validating instruments in MIS research. *Management Information Systems Quarterly, 13*(2), 147-169. https://doi.org/10.2307/248922

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209-231). New York, NY, USA: Guilford Publications.