Sum Scores Are Factor Scores

Daniel McNeish[1] & Melissa Gordon Wolf[2]

[1]Arizona State University, USA
[2]University of California, Santa Barbara, USA

Word Count: 6158

Contact Information:

Daniel McNeish, Department of Psychology, PO Box 871104, Tempe, AZ, USA 85287. Email: dmcneish@asu.edu

Abstract

Multiple-item scales are common in psychology whereby related items are administered and then a score for the overarching construct is created from the item responses. A common way to form this score is to sum responses of all items, as opposed to more statistically involved approaches such as factor analysis where the contribution of each item can vary. Though these methods are juxtaposed at times, this paper highlights how summing items is a special, highly constrained case of factor analysis. That is, when researchers sum score, they are still specifying the equivalent of a highly constrained factor model. This is relevant because reporting psychometric properties of sum scores is uncommon, making the psychometric properties of such scales unclear. Further, if researchers use factor analysis to validate a measure but subsequently sum score, they are employing a different, unvalidated model. By framing sum scoring as a type of factor analysis, our goal is to raise awareness that sum scoring is indeed a model that must be justified, like another other model. Validity, reliability, classification from sum scores cut-offs, and sum scoring previously validated scale are discussed to encourage researchers to more critically evaluate how they obtain and use scale scores in subsequent analyses.

Sum Scores Are Factor Scores

In psychological research, variables of interest frequently are not directly measurable (e.g., Joreskog & Sorbom, 1979). With constructs like motivation, mathematics ability, or anxiety, direct measures abate and the construct is instead captured via a set of items from which a single score (or small number of sub-scores) is calculated. Because these scales are not direct measures of the attribute (i.e., researchers cannot hold up a ruler to evaluate one's motivation), there is some ambiguity over how to create scores from these items. Such choices are not trivial and the flexibility possessed by the researcher can lead to scores that look quite different, even if scores materialize from the same data (e.g., Steegen et al., 2016). Such decisions have been considered to be an underemphasized source of replicability issues (Flake & Fried, 2019; Fried & Flake, 2018) as researcher degrees of freedom can be notable if not documented.

Several studies have reviewed the literature to inspect how researchers report the psychometric properties of the scales used in their studies (Barry et al., 2014; Crutzen & Peters, 2017; Flake, Pek, & Hehman, 2017) and the rigor that accompanies scales tends to be scant. For instance, Crutzen and Peters (2017) report that while nearly all in health psychology studies report some measure of reliability to accompany scale scores, less than 3% of studies in their review reported information about the validity of their scale – that is, whether the scale is measuring what it was intended to measure. When considering best practices in scale development, evidence for the internal structure of the scale is often recommended as a key component (e.g., Gerbing & Anderson, 1988). Assessment of internal structure is commonly done with latent variable models like factor analysis, which explore whether treating items as measurements of the same construct is supported empirically (Furr, 2011; Ziegler & Hagemann, 2015). However, as noted by Bauer and Curran (2015), it is much more common in psychology

to score scales by simply adding responses from multiple-item scales to create scores for variables that are not directly measureable rather than by performing a latent variable analysis. Flake et al. (2017) quantify this claim by reporting that 21% of studies reviewed using an established measure presented evidence of internal structure (37 out of 177 studies). Furthermore, just 2% of author-developed scales reported evidence of internal structure (3 out of 124). Combined, only 13% of studies provided evidence of validity based on the internal structure (40 out of 301 studies); an important source of evidence for multi-item scales (Standards for Educational and Psychological Assessment, 2014).

Sum scoring may be used for a variety of reasons such as to keep scale scores on an interpretable metric, to be consistent with previous studies, or avoid complex statistical modeling alternatives (e.g., DiStefano, Zhu, Mindrila, 2009; Hinz et al., 2012). Scale development recommendations in public health go as far to say, "In general, it does not make much difference in the performance of the scale if scales are computed as unweighted items (e.g., mean or sum scores) or weighted items (e.g., factor scores)." (Boateng et al., 2018, p. 13). The goal of this paper is to show why these arguments supporting sum scoring are not particularly persuasive and that sum scoring is an often unnoticed approach whose implementation and reporting practices often bear resemblance to other questionable research practices and one that could be contributing to the eroding reproducibility of psychological research.

As we will cover in this paper, sum scoring should not be considered an alternative to latent variable models because sum scoring *is a latent variable model*, albeit a highly constrained version. We start this paper by showing how sum scoring is *isomorphic* with a parallel factor model, meaning the two methods yield results that are a perfect linear transformation of one another. We discuss how sum scoring makes it easy to avoid reporting psychometric properties

of scales – evidence for internal structure, in particular – whereas psychometric properties can be evaluated far more easily with factor models. We discuss ramifications for converting quantitative scale scores to qualitative classification groups with sum scoring in addition to the widespread practice of sum scoring previously validated models, even though sum scoring in this way changes the model and renders previously reported evidence moot. We emphasize this point to engage readers who believe that latent variable modeling is only necessary for validation, or that using a "previously validated" scale alleviates the need to use a factor model in subsequent analyses or inferences.

Most importantly, we argue that sum scoring *is* a factor model and the two approaches should be reported identically rather than juxtaposed. We do not necessarily argue that sum scoring are *always* detrimental but rather that *justification* for sum scoring and reporting of supporting evidence is often lacking because the approach appears arithmetic and model-free when, in fact, it is isomorphic with a highly restricted, heavily constrained parallel factor model. Our ultimate goal is to convince researchers that scoring scales – by any method – is a statistical procedure that requires evidence and justification– regardless of method – as would any other model.

Sum Scoring as a Parallel Factor Model

Consider 6 items from 301 students aiming to measure cognitive ability from the classic Holzinger and Swineford (1939) data, which are publically available from the `lavaan` R package (Rosseel, 2012) [all data, results, and analysis code are available on the Open Science Framework, https://osf.io/cahtb/). The item scores range from 0 to 10; some of the original items contain decimals, but we have rounded all items to the nearest integer. Table 1 shows a brief description of each of these items along with basic descriptive statistics.

Table 1
Item descriptions and item descriptive statistics

Item	Description	Mean	Std. Dev	Min	Max
1	Paragraph Comprehension	3.09	1.17	0	6
2	Sentence Completion	4.46	1.33	1	7
3	Word Definitions	2.20	1.13	0	6
4	Speeded Addition	4.20	1.15	1	7
5	Speeded Dot Counting	5.56	1.03	3	10
6	Discrimination Between Curved and Straight Letters	5.37	1.08	3	9

To sum score these 6 items, the scores of each item would simply be added together,

$$\text{SumScore} = \text{Item}1 + \text{Item}2 + \text{Item}3 + \text{Item}4 + \text{Item}5 + \text{Item}6 \quad (1)$$

Sum scores *unit-weight* each item (Wainer & Thissen, 1976), meaning that we could equivalently write Equation 1 with a "1" coefficient (or any other arbitrary value so long as it is constant) in from of each item,

$$\text{SumScore} = 1\times\text{Item}1 + 1\times\text{Item}2 + 1\times\text{Item}3 + 1\times\text{Item}4 + 1\times\text{Item}5 + 1\times\text{Item}6 \quad (2)$$

Unit-weighting is possible in factor models by constraining all standardized loadings to the same value. In psychometric terms, this is referred to as a *parallel model* such that the unstandardized loadings and error variances are assumed identical across items (Graham, 2006). In the factor model context, the true score of the construct under investigation is modeled as a latent variable, which predicts the item scores for each item.[1] This means that the observed item

[1] There is a deep literature on the differences between *reflective* latent variables and *formative* latent variables (e.g. Bollen, 2002; Bollen & Lennox, 1991; Borsboom, Mellenbergh, & van Heerden, 2003; Edwards & Bagozzi, 2000). The sum score formulation in Equation 1 might be more closely viewed as formative latent variable where the observed indicators are the predictors and the latent variable is the outcome, rather than the reflective model shown in Equation 3 where the observed indicators are the outcome and the latent variable is the predictor. We concede these nuances and continue with reflective latent variable models given that (a) the goal of this paper is to generally show researchers that sum scores should be considered latent variables, (b) the decision between a formative and reflective latent variable is often subjective and varies on a case-by-case basis (Bagozzi & Edwards, 2000), and (c) reflective models are by far more common in psychology (Bollen, 2002) and appear to be a more desirable starting point to accomplish (a). It has also been noted that the two different specifications often lead to the same results, practically (e.g., Goldberg & Digman, 1984; Fava & Velicer, 1992; Reise, Waller, & Comrey, 2000).

values are the outcome on the left-hand side of the equation rather than being a variable on the right-hand side. The factor model equations would therefore be

$$\begin{aligned}
\text{Item}1_i &= \tau_1 + \lambda_1\zeta_i + \varepsilon_{1i} \\
\text{Item}2_i &= \tau_2 + \lambda_2\zeta_i + \varepsilon_{2i} \\
\text{Item}3_i &= \tau_3 + \lambda_3\zeta_i + \varepsilon_{3i} \\
\text{Item}4_i &= \tau_4 + \lambda_4\zeta_i + \varepsilon_{4i} \\
\text{Item}5_i &= \tau_5 + \lambda_5\zeta_i + \varepsilon_{5i} \\
\text{Item}6_i &= \tau_6 + \lambda_6\zeta_i + \varepsilon_{6i}
\end{aligned} \tag{3}$$

where τ is the item intercept (the average item response score across all people), λ is the loading from the latent true score ζ for person *i* to the observed item score (the regression coefficient from the true score to the observed item response), and ε is an error term for person *i* (the difference between the predicted item response and the observed item response). The latent variable and the errors are assumed to be normally distributed with a mean of 0 and an estimated variance: $\zeta \sim N(0, \psi)$ and $\boldsymbol{\varepsilon} \sim MVN(\mathbf{0}, \boldsymbol{\Theta})$. To extend Equation 2 in a factor model context, a parallel model can be specified by constraining all the λ values in Equation 3 to be the same value and all the diagonal values of $\boldsymbol{\Theta}$ are set to be equal and all off-diagonal terms are set to be zero (i.e., a homogeneous conditional independence structure). This is the factor model that is applied when one uses a sum score.

To make this model clearer, the path diagram for a parallel model is shown in Figure 1. The 1.0s on the factor loadings indicate that the loadings are constrained to be equal and the θ value on each of the residual variance indicate that these values are all constrained to be equal. The loadings need not be constrained to 1.0, but they all need to be constrained to the same value. Not shown are the estimated item intercepts for each item (the τ parameters); estimating the intercept for each items results in a saturated mean structure so that the item means are just

equal to the descriptive means of each item (assuming no missing data). The mean of the latent factor is constrained to 0 as a result.

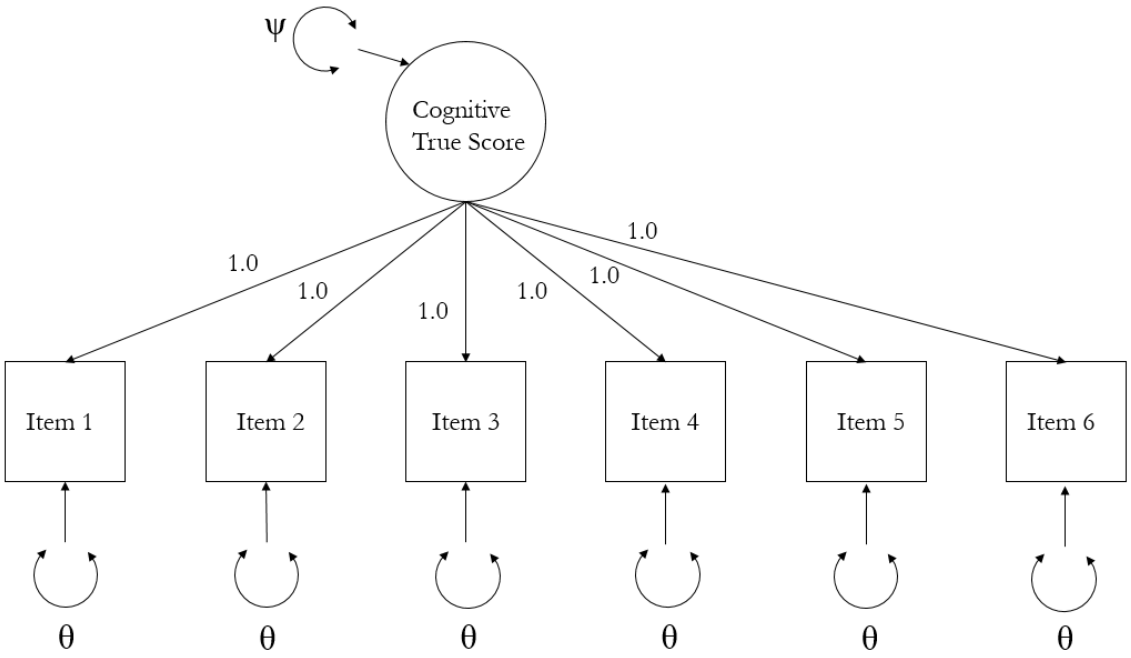

Figure 1. Path diagram of parallel measurement model that unit weights items as a factor analysis model. The residual variance is estimated but constrained to be equal for all items. Each of the loadings are constrained to 1 for all items. The latent variable variance is estimated. Intercepts for each item are included but are not shown. The latent variable intercept is constrained to 0.

We fit this parallel model to these 6 cognitive ability items in M*plus* Version 8.2 with maximum likelihood estimation and saved the estimated parallel model factor scores for each person. We then compared the parallel model factor scores to the sum scores. The scatterplot with a fitted regression line for this comparisons is shown in Figure 2. Notably, the R^2 for the regression of the factor scores on the sum scores is exactly 1.00 (meaning that the correlation between the two is also 1.00). Depending on how the model is parameterized, the factor scores

from the parallel model with not be exactly equal to the sum scores;[2] however, there will necessarily be a perfect linear transformation from parallel model factor scores to sum scores under any parameterization of the parallel model.

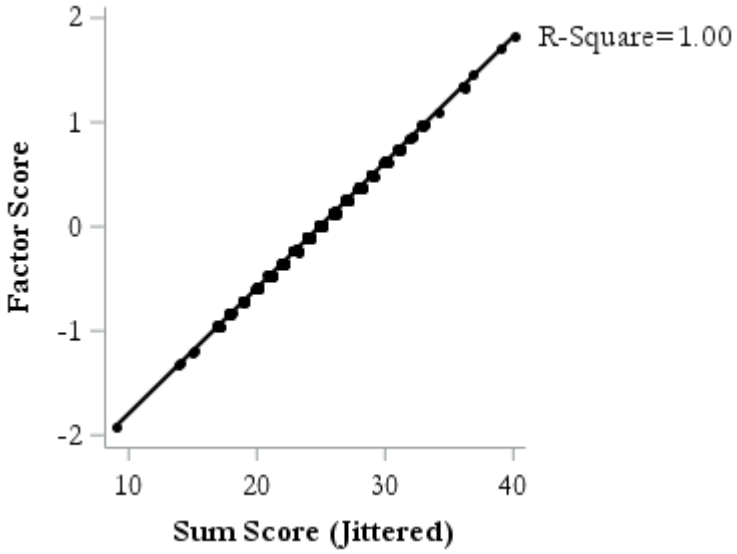

Figure 2. Jittered scatter plot of sum scores with factor scores from the model in Figure 1 with a fitted regression line. N =301

An Alternative to the Parallel Model: The Congeneric Model

Whereas sum scoring can be expressed (through a linear transformation) as a parallel factor model, *optimal weighting* of items with a *congeneric model* is a more general approach. The basic idea of a congeneric model is that every item is differentially related to the construct of interest and every item has a unique variance (Graham, 2006). So if Item 1 is more closely related to the construct being measured that Item 4, Item 1 receives a higher loading than Item 4. Conceptually, this would be like having different coefficients in front of each item in Equation 2 so that each item is allowed to more strongly or more weakly correspond to the construct of

[2] Rose, Wagner, Mayer, & Nagengast (2019) show the constraints that need to be imposed to yield the exact sum scores. Given the complexity required to achieve this, we proceed with the simpler approach that yields a perfect linear transformation but not the exact sum score, which remains sufficient to make our desired points.

interest. In the factor model, this would mean that the each λ loading could be estimated as a different value and that each residual variance θ would be uniquely estimated as well (i.e., the latent variable accounts for a different amount of variance in each item).

Figure 3 shows the path diagram of a congeneric model for the same data used in Figure 1. The major difference is that the loadings from the latent variable to each observed item are now uniquely estimated for each item, as are the residual variances for each item (noted by the subscripts on the parameter labels in Greek letters). In order to uniquely estimate the loadings for each item, the variance of the latent variable is constrained to a value (1.0 is a popular value to give this latent variable a standardized metric).

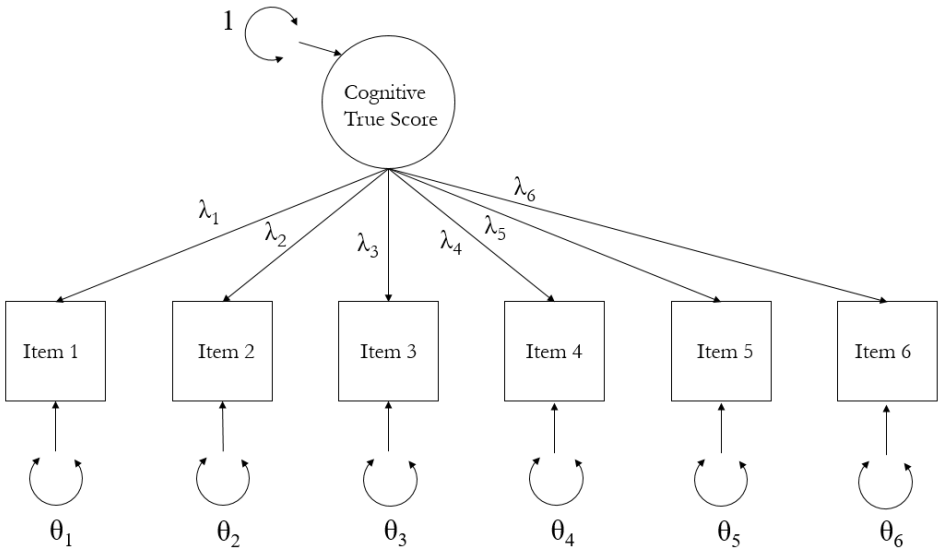

Figure 3. Path diagram of congeneric factor analysis model. The residual variance is uniquely estimated for each item as are the loadings for each item. The latent variable is given scale by constraining its variance to 1.0. If the latent variable variance were of interest, scale could alternatively be assigned by constraining one of the loadings to 1. Intercepts for each item are included but are not shown. The latent variable intercept is constrained to 0.

We fit a congeneric model to the six cognitive ability items in M*plus* Version 8.2 with maximum likelihood estimation and saved the estimated congeneric model factor scores for each person. The standardized loadings, unstandardized loadings, and residual variances from this

model are shown in Table 2. Of note is that the standardized loadings are quite different across the items in Table 2, suggesting that each item relates to the latent variable quite differently and that it would likely be inappropriate to unit-weight these items.

Table 2
Model estimates from congeneric model in Figure 3

Item	Description	Std. Loading	Unstd. Loading	Residual Variance
1	Paragraph Comprehension	0.82	0.96	0.44
2	Sentence Completion	0.85	1.12	0.50
3	Word Definitions	0.79	0.89	0.47
4	Speeded Addition	0.17	0.20	1.28
5	Speeded Dot Counting	0.18	0.19	1.02
6	Discrimination between curved and straight letters	0.26	0.28	0.11

Figure 4 shows the scatterplot and fitted regression line for sum scores against the congenic model factor scores. Notably, the R^2 value is 0.76 and the two scoring methods are far from identical, unlike the relation between sum scores and parallel model factor scores. This means that two people with an identical sum score could have potentially different congeneric model factor scores because they reached their particular sum score by endorsing different items. Because the congeneric model weights items differently, each item contributes differently to the factor score (but not to the sum scores). Therefore, the congeneric model factor scores are taking into account not just *how* an individual responded to each item, but also to *which* items these responses occur.

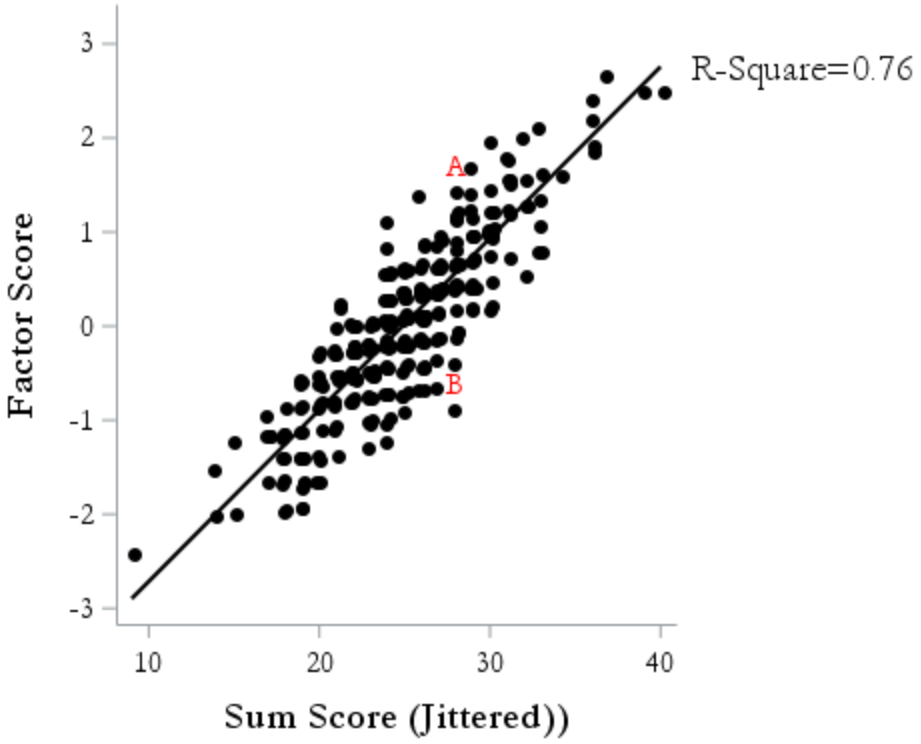

Figure 4. Jittered scatter plot of sum scores with congeneric factor scores from the model in Figure 3 with a fitted regression line. Labels A and B are used to demonstrate a point later on in the "Classification" Section. Data labels appear directly above the relevant data point. *N* =301

Importance for Psychometrics: Reliability Coefficients

Though the isomorphism between sum scores and parallel model factor scores may seem little more than a statistical sleight of hand, the equivalence can be important for judging psychometric properties of multiple item scales. Reliability is the most frequently reported psychometric property of scales in psychology (e.g., Dima, 2018). By far, the most popular metric for reliability is Cronbach's alpha (Hogan, Benjamin, & Brezinski, 2000). However, as methodologists have noted (e.g., Dunn, Baguley, & Brunsden, 2014; Green & Yang, 2009; McNeish, 2018; Zinbarg, Yovel, Rvelle, & McDonald, 2006), Cronbach's alpha makes an assumption of *tau-equivalence*, which is similar to the parallel model in that it assumes equal loadings across all items, though tau-equivalence relaxes the assumption that the variances are

also equal (Graham, 2006). When a scale is truly congeneric (which is often more common than tau-equivalence in practice; Peterson & Kim, 2013), alpha underestimates reliability and the degree of underestimation increases as the scale increasingly diverges from tau-equivalence (Graham, 2006; Miller, 1995; Yang & Green, 2011).

When scales are congeneric, more general measures of reliability that do not assume tau-equivalence tend to more appropriate (Peters, 2014; McNeish, 2018; Revelle & Zinbarg, 2009; Sijtsma, 2009) such as coefficient *H* developed for optimally-weighted scales (Hancock & Mueller, 2001). This pattern can be seen with the Holzinger and Swineford (1939) cognitive ability data. If assuming that the scale can be unit-weighted, the Cronbach's alpha estimate of reliability is 0.72. If using a congeneric model and concluding that the scale should be optimal-weighted, the estimate of reliability from coefficient *H* is 0.87. Because the standardized loadings for the different items vary considerably in this data (range: .17 to .85), there is a sizeable difference between the different reliability estimates given the difference in their assumptions and given the clear violation of tau-equivalence. Therefore, sum scoring items ignores possible differences in the contribution of each item, which could lead to researchers reporting unnecessarily low reliability values for their scales (possibility increasing file drawer problems since measures with low reliability are at higher risk in the review process).

Importance for Psychometrics: Classification

In some areas of psychology, cut-offs are applied to quantitative scales to create meaningful, qualitatively distinct groups. This is especially common in clinical psychology with scales like Beck's Depression Inventory (BDI), the PTSD Checklist (PCL-5), the Hamilton Depression Rating Scale, and the State-Trait Anxiety Inventory, among others. Each of these scales can be scored using a sum score, which can be used to classify participants into clinical

groups. For example, depression is classified from the BDI as “Minimal” for sum scores below 13, “Mild” for scores from 14 to 19, “Moderate” for scores from 20 and 28, and “Severe” for scores from 29 to 63 (Beck, Steer, & Brown, 1996). Though we recognize the use of this and related inventories in clinical settings as a quick, simple approximation, such a use is harder to defend in rigorous research studies (e.g., when the scales are used as outcome measures to determine the efficacy of treatment). With clinical scales that often include many items (e.g., the BDI contains 21 items), the assumption of sum scoring that all items measure features of the construct equally becomes less plausible. If all items do not contribute equally to the construct, then it matters *which* items are strongly endorsed, not necessarily just *how* many items were strongly endorsed. In other words, an item about suicidality might warrant more attention than an item about a common side effect from treatment, such as weight gain or troubling sleeping.

Consider again the case of the Holzinger and Swineford (1939) data. In this data, the loadings of each of the items are quite different, so students with the same sum scores can end up with different congeneric model factor scores depending on the response pattern than yielded the sum score. For instance, consider Student A whose 6 item responses for Item 1 through Item 6 (respectively) were (5, 6, 4, 3, 5, 5) and Student B whose respective responses were (2, 3, 1, 5, 10, 7). The sum score of both students is 28 but the congeneric model factor scores are wildly different because the loadings of Items 4 through 6 were low, indicating that these items are weakly related to the latent variable. Because Student B (labeled as “B” in Figure 4) scored poorly on the most meaningful items for the latent variable (Items 1 through 3), their congeneric model factor score was estimated to be -0.88 (the factor score is on a *Z*-scale, so this is well below average). Conversely, the Student A’s congeneric model factor score was estimated to be

1.43 (well above average) given that they were near the sample maximum score for the first three items (Student A's data is labeled as "A" in Figure 4).

Even though sum scores would consider these students to be the same, the congeneric model factor scores indicate that their ability is disparate. The congeneric factor model was parameterized such that the factor scores were from a standard normal distribution, meaning a sum score of 28 corresponds to about 74% of the distribution (the area between a Z-score of -.88 and a Z-score of 1.43), an extremely large range that shows the potential imprecision of unit-weighting when it is inappropriate. Thus, "cutting" the sum score could result in classifying people with dramatically different factor scores in the same category.

Importance for Psychometrics: Metrics of Validity

When multiple items are summed to form a single score, it is difficult and therefore uncommon to report on the internal structure of the scale (Crutzen & Peters, 2017). However, as mentioned earlier, sum scores are a perfect linear transformation of factor scores from a parallel factor model. By representing sum scoring through parallel model factor scores, researchers can present evidence from the fit measures in the factor model framework to determine whether sum scoring is a reasonable approach. Though arguments continue in the statistical literature about the best way to assess model fit for factor models (e.g., Barrett, 2007; Milsap, 2007; Mulaik, 2007), popular options include fit statistics (e.g., the T_{ML} statistic; a.k.a. the χ^2 test) or approximate goodness of fit indices (e.g., SRMR, RMSEA, or CFI).

For the parallel model fit to the Holzinger and Swineford (1939) data in Figure 1, model fit is quite poor by essentially any metric.

- The CFI value is 0.45 whereas values at or above 0.95 are considered to indicate good fit (e.g., Hu & Bentler, 1999).

- The SRMR is 0.24 which does not compare favorably to the 0.08 or lower cut-off that is commonly recommended.
- The RMSEA value is 0.23 (90% CI = [0.21, 0.25]), which similarly exceeds the recommendation for good fit of 0.06 or lower.
- The maximum likelihood test statistic (T_{ML}) is also significant, $\chi^2(19) = 5361.86, p < .001$ which suggests that the model-implied mean and covariance structures differ from those from the data.

Taken together, these tests of model fit clearly show that the parallel model (and the linear transformation of sum scores it produces) is not supported empirically.

Next, we test the fit of the congeneric model from Figure 3. The fit of this model is not great either – CFI = 0.81, SRMR = 0.11, RMSEA = 0.20 [90% CI = (0.17, 0.23)], and $\chi^2(9) = 115.37, p < .001$. Although the fit improved, the values are still not in the acceptable range for any of the measures here. This example shows a benefit of considering scales in the factor model framework: we know now that these items are not measuring the same thing and that there may be multiple constructs being measured by these six items, an aspect that is easy to overlook with sum scores. Seeing the poor fit of the one-factor congeneric model and the disparate loadings in Table 2, it seems like the first three items are more related to verbal skills whereas the second set of three items are more related to speeded tasks. Therefore, we fit a two-factor model where Items 1 through 3 load on one factor and Items 4 through 6 load on a second factor, with the factors being allowed to covary. The path diagram with estimated standardized loadings and the estimated factor correlation is shown in Figure 5. The fit of this model is much improved – CFI = 0.99, SRMR = 0.03, RMSEA = 0.05 [90% CI = (0.00, 0.10)], and $\chi^2(8) = 14.74, p = .07$, providing empirical support for this internal structure of the scale.

Practically speaking, summing all items in a scale that has been validated using a multi-factor model results in incorrect interpretations.

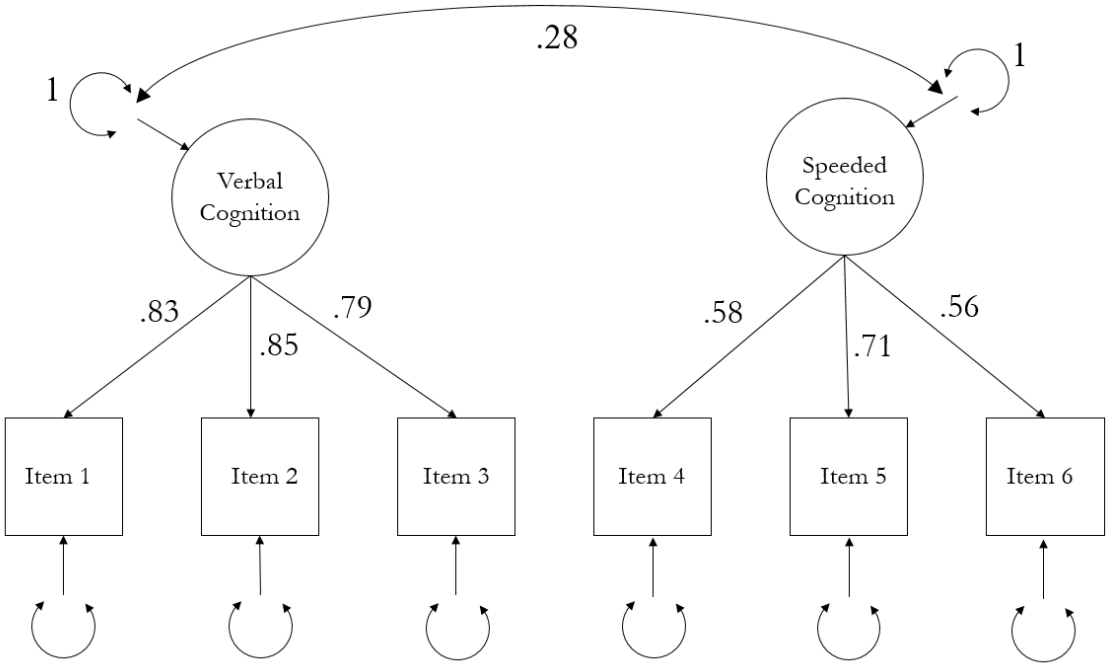

Figure 5. Path diagram of two-factor congeneric model with standardized factor loading estimates and estimated factor correlation

Importance to Psychometrics: Previously Validated Scales

Scales that are widely used in practice are often accompanied by a citation to a validation study providing evidence for the internal structure and the reliability of the scale. In many cases, these studies are performed using some type of factor model. However, when many of these validated scales are used in practice, scores are derived from summing the items, despite the fact that validation studies routinely fit congeneric models with different loadings for each of the items (Houts & Edwards, 2019). Alluding to our previous point, the issue here is that sum scoring is a factor model but it is not the same factor model that was used to validate the scale. Validation studies provide evidence of the internal structure under a congeneric model but if the scoring model then reverts to a sum score, the validation study is no longer applicable as

evidence. In this scenario, the model used for validation (congeneric) and the model used for scoring (a linear transformation of a parallel model) are incongruent and new evidence would be required to confirm that the same internal structure holds with a parallel model to justify sum scoring the scale.

As a quick example, we revisit two scales discussed earlier: the Beck Depression Inventory (BDI) and the PTSD Checklist (PCL-5). The BDI in particular is a high stakes assessment since it is often used as an outcome metric in clinical depression trials (Santor, Gregus, & Welch, 2009). As mentioned earlier, the BDI is scored using a sum score across all items (per the BDI manual; Beck, Steer, & Brown, 1996) and participants are classified into qualitatively meaningful groups using cut scores. The PCL-5 can be scored three ways: (a) by summing all items, (b) by summing items within a cluster, or (c) by counting the number of times items have been endorsed within each cluster (Weathers, et. al., 2013). There are different cut scores associated with each scoring method.

The primary BDI validation paper (Beck, Steer, & Garbin, 1988) has been cited 12,330 times and the primary PCL-5 validation paper (Blevins, et. al., 2015) has been cited 691 times on Google Scholar at the time of this writing. In these papers, the BDI was validated as a two-factor congeneric model while the PCL-5 was validated as either a four-factor or six-factor congeneric model. Notably, neither of these validated psychometric models align with the model that corresponds to the recommended scoring methods; the scales are scored using a completely different factor model (i.e., summing across all items implies the use of a unidimensional parallel factor model) compared to the model used for validation (i.e., a multi-dimensional congeneric factor model). In other words, in their current uses, the BDI and the PCL-5 have not demonstrated quantitative psychometric evidence of validity based on the internal structure (at

least, within their respective top cited validation publications) despite many empirical studies suggesting otherwise.

Our intention is not to single out these two scales as sum scoring is a common practice whose correspondence to highly constrained factor models is not appreciated. However, as noted by Fried and Nesse (2015), creating unidimensional sum scores for multi-dimensional constructs may obfuscate findings in psychological research. When assessments are scored differently, utilize cut scores, and do not align with the validated model, it can be difficult to find meaningful, consistent results across studies.

Furthermore, best practice recommendations call for scale validation to be carried out each time the scale is used because validity is not a property of the scale itself but instead concerns interpretation of each isolated use (Kane, 2006). That is, previous validation of a scale is not sufficient evidence that the scale is valid for how it is being used in any single context. This recommendation is of particular relevance in the case of sum scoring previously validated scales: if the parallel model used to obtain (a linear transformation of) sum scores were assessed for model fit, the results may look much different (and likely worse) than the model used in validation studies cited as evidence for the internal structure of the scale. This returns to the central thesis of the argument in this paper – sum scoring is a model and sum scoring a previously validated (congeneric) scale is inherently fitting a different model with different properties than the model reported in the validation study.

Are Sum Scores Ever OK?

Consider the two-factor congeneric model from the Holzinger and Swineford (1939) data presented earlier. We showed that the scale far more plausibly represented two distinct constructs (Verbal and Speeded Cognition) based on the model fit assessment. Recall from Figure 6 that the

standardized factor loadings in were very close for the Verbal factor (.83, .85, .79) and the standardized loadings for the Speeded Cognition scales did not deviate too greatly (.58, .71, .56), which may mean that assumption violations of the parallel model may be minimal and sum scoring will not produce noticeably different result from congeneric model factor scores.

We fit a two-factor parallel model to these data in M*plus* 8.2. The loadings for all items were constrained to 1.0 and the residual variances were constrained to be equal across all items within each factor but were uniquely estimated across factors. The factor variances were also uniquely estimated but factors were not allowed to covary in order to retain isomorphism between the parallel model factor scores and sum scores for each subscale. The path diagram for this two-factor parallel model is shown in Figure 6.

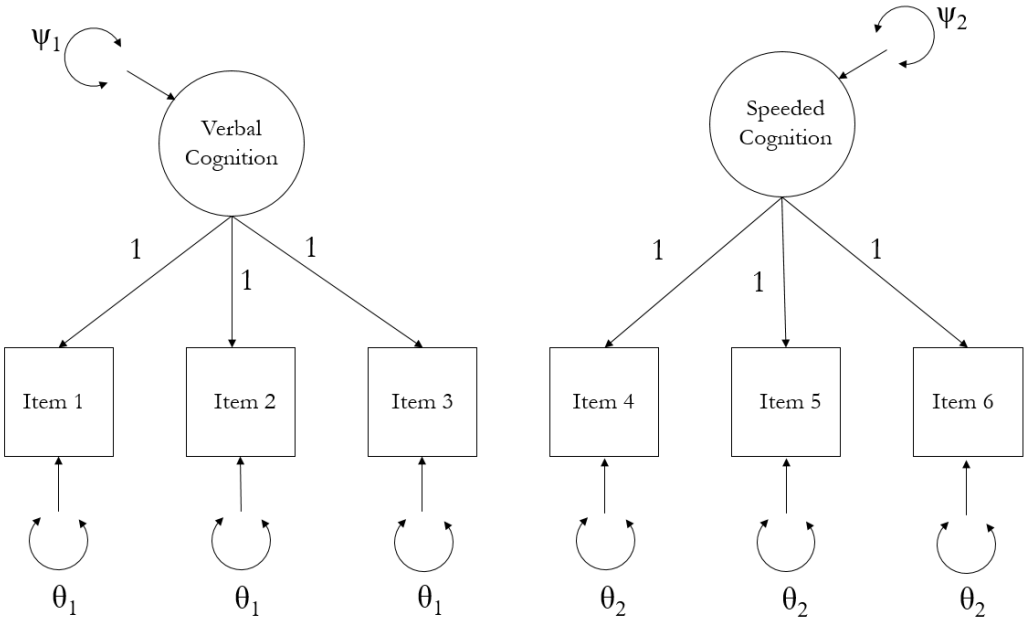

Figure 6. Path diagram of two-factor parallel model. The loadings are constrained to 1 for all items, the residual variances are unique across factors but are constrained within factors. Factor variances are uniquely estimated and there is no factor covariance. Intercepts for each item are included but are not shown. The latent variable intercepts are constrained to 0 for each factors.

First, Figure 7 shows the correlation between the two-factor parallel model factor scores and the sum scores. As shown above and as expected, the parallel model yields factor scores that are a perfect linear transformation of the sum scores and the correlation is exactly 1.00. Second, we inspected the fit of the parallel model: CFI = 0.93, SRMR = 0.14, RMSEA = 0.09 [90% CI = (0.06, 0.11)], and $\chi^2(17) = 55.54$, $p < .01$. The fit of the model is not great, but might be interpreted to show some weak indications of good fit (e.g., a CFI above .90 is sometimes considered sufficient, 90% CI of RMSEA contains .06). A likelihood ratio test comparing the two-factor parallel model to the two-factor congeneric model from Figure 5 shows that the congeneric model fits significantly better, $\chi^2(9) = 40.80$, $p < .01$.

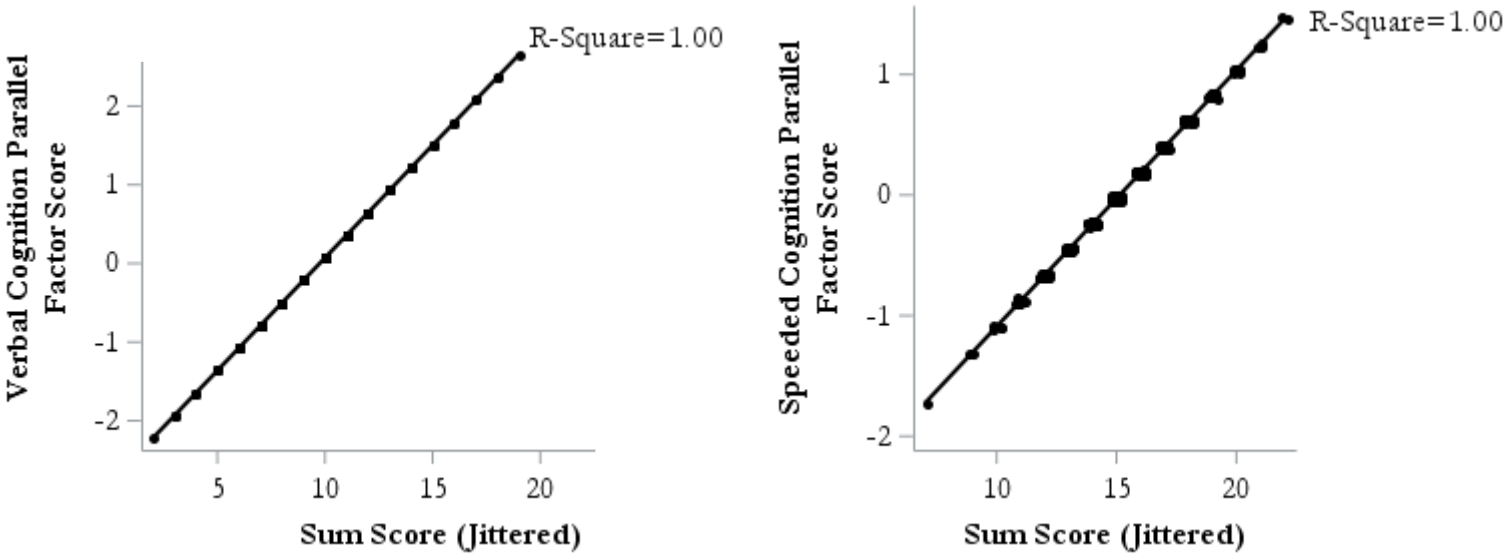

Figure 7. Jittered scatter plot of sum scores with parallel model factor scores from the model in Figure 6, with a fitted regression line. Verbal Cognition is shown in the left panel and Speeded Cognition is shown in the right panel. *N* =301

However, if the sum scores are directly compared to the factor scores from the congeneric model, the R^2 values are quite high: 0.99 for the Verbal Cognition factor and 0.96 for the Speeded Cognition factor (keep in mind that there only three items per factor in this example;

the inclusion of additional items gives more opportunity for loadings to vary across items). These relations are plotted in Figure 8. The extremely close standardized loadings for the Verbal Cognition factor led to sum scores that are almost identical to the congeneric factor scores; the standardized loadings for the Speeded Cognition factor are more distant, so the differences are easier to detect (also note that even at a R^2 of .96, the range of congeneric factor scores within each sum score remains about half a standard deviation; this discrepancy would be problematic in a high-stakes contexts). In general, the larger the differences are in the standardized loadings are for items that load on the same factor, the larger the differences will be between sum scores and congeneric model factor scores (Wainer, 1976). When the standardized loadings are very close for items that load on the same factor, there will be little difference between sum scores and congeneric factor scores.

Additionally, there is not much difference in the reliability of the scale based on the scoring method; alpha on the sum scores was .86 for the Verbal Cognition factor and .64 for the Speeded Cognition factor whereas Coefficient *H* was .87 for the Verbal Cognition factor and .66 for the Speeded Cognition factor. Though the different approaches vary with respect to how well they fit the data, the strong correlations suggest that scores – especially for Verbal Cognition scores – are less sensitive to choice of scoring method (sum scoring or congeneric factor scoring). To reiterate, this is true *only* when items have similar loadings on the *same* factor. Though we would still recommend that the congeneric scores be used because it is still better able to detect nuance and its overall fit is significantly improved, one could construct an argument for sum scoring each subscale (i.e., items on each factor) in this data if there is some preferable interpretation based upon sum scores, understanding the risks associated with cut-

scores and high-stakes contexts (i.e., incorrectly classifying persons or evaluating treatment efficacy in clinical studies).

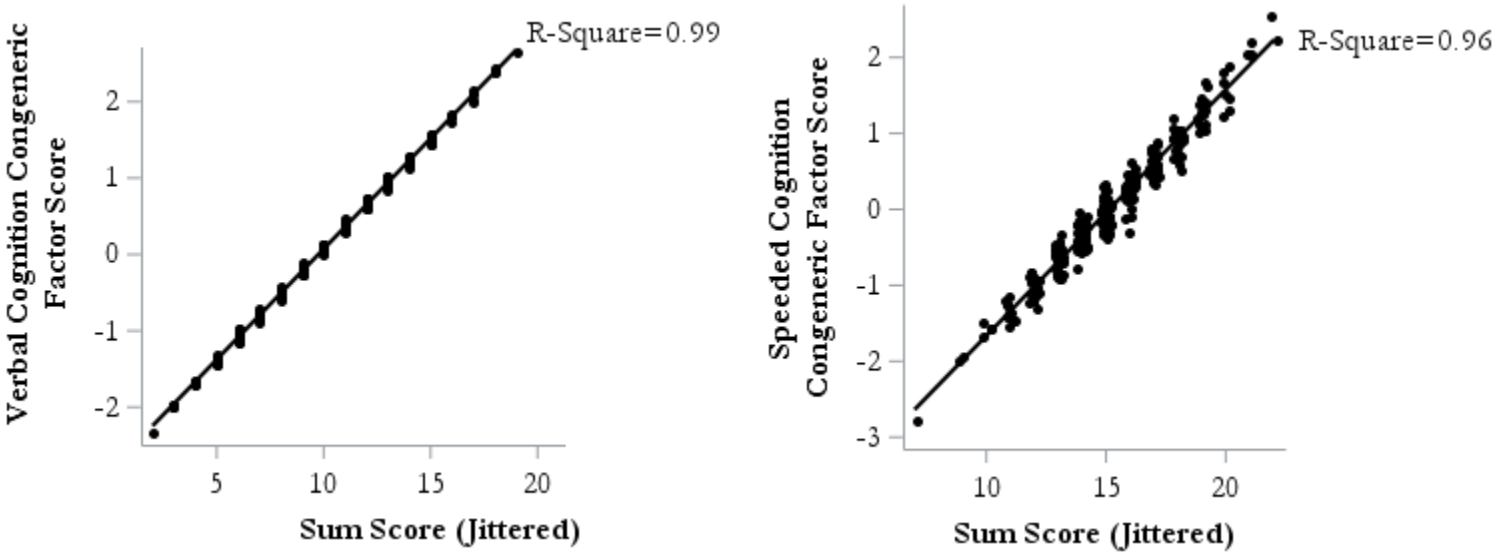

Figure 8. Jittered scatter plot of sum scores with congeneric factor scores from the model in Figure 5, with a fitted regression line. Verbal Cognition is shown in the left panel and Speeded Cognition is shown in the right panel. *N* =301

Discussion & Limitations

Given the nature of the topics under investigation in psychological science, many research studies rely on multiple-item scales to tap constructs that are not directly measureable with physical instruments. Variables created from scoring these scales often play a central role in subsequent analyses, either as focal predictor variables or as the primary outcome of interest. However, when justification for the scoring of scales is relegated to secondary status as is often the case when sum scores are created, it can lead to hidden ambiguity in research conclusions about the intrinsic meaning represented by the variable. For instance, if we had items from a Five Factor personality scale and did not consider the internal structure of the items and simply

summed all the responses across all items, how would the score be interpreted? What would it mean to have average personality?

The scores from multiple-item scales are treated seriously but the process by which those scores are obtained typically is not. There are countless modeling options that one can make that lead to the creation of these scores: (a) are the items treated as continuous or discrete, (b) do any response categories need to be collapsed or reverse coded, (c) are there subscales present in the scale? Whenever responses from multiple items are combined by some method, there is a model corresponding to that method. Although summing item responses may seem like a simple arithmetic operation, it is a simple linear transformation of a parallel factor model.

Our point is that any method that is advanced by researchers for scoring scales needs evidence to support its use. The field would not endorse conclusions that were not supported by statistical evidence in some form (e.g., *p*-values, Bayes factors, variance explained measures, etc.), so why does the field so readily accept conclusions derived from scoring multiple-item scale scores without any accompanying evidence? Such *v*-hacking and *v*-ignorance (where *v* is shorthand for validity; Hussey & Hughes, 2019) may be at the foundation of the replication and measurement issues in psychology; if scales are scored using untested psychometric models with unknown or questionable properties, it is difficult to replicate findings or infer meaning.

The goal here is not to disparage sum scoring or whatever other scoring method a researcher might be interested in using. Rather, we are trying to make the point that *all* of these choices are models and *any* choice should be accompanied by evidence. Sum scoring is not a particularly complex model, but it is still a model nonetheless. If a researcher wants to create scale scores by summing item responses together, that is not necessarily deficient in all cases, as alluded to above. However, some evidence needs to be reported to support that decision: is the

internal structure supported? Are loadings sufficiently similar such that each of the items contribute about equally to what is being measured? Is there a discernible difference in scores from a parallel model and scores from a congeneric model? Are there changes in reliability of the scores with different scoring methods? If a researcher can justify sum scoring (or a linear transformation thereof via a parallel model), then sum scores are not an inherent problem. Some methodologists may be diametrically opposed to sum scores in any context, but arguments emanating from differing thought processes are a natural part of scientific debate. The real problem that permeates throughout the field is employing methods without any justification. So while we will not advise one particular method over any other, we do implore researchers to take psychometrics as seriously as other statistical procedures and provide justification for whichever method they choose.

Limitations

First, contrary to common perception in the field, cut-offs for model assessment measures for factor models are not definitive. The commonly reference Hu and Bentler (1999) cut-offs were based on empirical simulation rather than analytic derivation and therefore are limited by the conditions included in the simulation design. Several studies have noted that the cut-offs for many popular indices – including CFI, RMSEA, and SRMR that we use in this paper – vary with the size of the loadings (Hancock & Mueller, 2011; McNeish, An, & Hancock, 2018), size of residual variances (Heene, Hilbert, Draxler, Ziegler, & Buhner, 2011), model type (Fan & Sivo, 2005), model size (Shi, Lee, & Terry, 2018), degree of misspecification (Marsh, Hau, & Wen, 2004), and missing data percentage (Fitzgerald, Estabrook, Martin, Brandmaier, & von Oertzen, 2018). We openly acknowledge the lack of firm recommendations on how to adjudicate what constitutes a “good” fitting model, but ultimately believe that imprecise metrics can be better

than no metrics at all. Additionally, latent variable fit is but one component and readers who are hesitant to rely on fit metrics may alternatively use correlations between parallel and congenic models, changes in reliability, or similarity of factor loadings.

Second, this only addresses one kind of evidence of validity (evidence based on the internal structure) and one quantitative method that could be used to provide such evidence (factor analysis). The *Standards for Educational and Psychological Assessment* name five types of evidence, none of which are inherently more important than the other. There is an extensive literature on the theory of measurement itself; for example, Maul (2017) demonstrates that good fitting models are not inherently evidence of good theory, Borsboom, Mellenbergh, & van Heerden (2004) discredit the nomological network and argue that validity is simply the causal relationship between variation in the attribute and variation in the response, while Michell (2012) argues that measurement is not possible in the social sciences as social scientists have not established evidence of quantitivity in the attributes they claim to measure. For this reason, we focused on classic, widely reported quantitative methods such as Cronbach's alpha and factor analysis since their misuse may represent a core component of replicability issues facing psychologists. Variables are the foundation of any statistical analysis, but methodological principles devised to combat replicability issues are irrelevant if the foundational unit to which they are applied is ultimately meaningless. We offer this paper as a starting point to hopefully bridge readers from reflexively sum scoring to the technical literature on scales and psychological measurement.

Author Contributions

Both authors jointly generated the idea for the paper. D. McNeish selected the data, analyzed the data, and created the figures. Both authors took part in writing the first draft. Both authors critically edited subsequent drafts and both authored approved of the final version for submission.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*, 815-824.

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability reporting practices in the field of health education and behavior. *Health Education & Behavior*, *41*, 12–18.

Bauer, D.J. & Curran, P.J. (2015). The discrepancy between measurement and modeling in longitudinal data analysis. In J.R. Harring, L.M. Stapleton & S.N. Beretvas (Eds.), *Advances in Multilevel Modeling for Educational Research* (pp. 3-38). Information Age Publishing.

Beck, A.T., Steer, R.A., & Brown, G.K. (1996). Manual for the Beck Depression Inventory-II. San Antonio, TX: Psychological Corporation.

Beck, A. T., Steer, R. A., & Garbin, M. G. (1988). Psychometric properties of the beck depression inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, *8*, 77–100.

Blevins, C. A., Weathers, F. W., Davis, M. T., Witte, T. K., & Domino, J. L. (2015). The posttraumatic stress disorder checklist for DSM-5 (PCL-5): Development and initial psychometric evaluation. *Journal of Traumatic Stress*, 28, 489–498. https://doi.org/10.1002/jts.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in Public Health, 6,* Article 149 (pp. 1 -18).

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*, 605-634.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305-314.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203-219.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

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

Dima, A. L. (2018). Scale validation in applied health research: Tutorial for a 6-step R-based psychometrics protocol. *Health Psychology and Behavioral Medicine*, *6*, 136-161.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*, 1-11.

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

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*, 155-174.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling*, *12*, 343-367.

Fava, J. L., & Velicer, W. F. (1992). An empirical comparison of factor, image, component, and scale scores. *Multivariate Behavioral Research*, *27*, 301-322.

Fitzgerald, C. E., Estabrook, R., Martin, D. P., Brandmaier, A. M., & von Oertzen, T. (2018, May 8). Correcting the bias of the root mean squared error of approximation under missing data. https://doi.org/10.31234/osf.io/8etxa

Flake, J. K., & Fried, E. I. (2019). Measurement schmeasurement: Questionable measurement practices and how to avoid them. PsyArxiv preprint, https://doi.org/10.31234/osf.io/hs7wm

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*, 370-378.

Fried, E. I., & Flake, J. K. (2018). Measurement matters. *APS Observer*, *31*.

Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*, 1–11. https://doi.org/10.1186/s12916-015-0325-4

Furr, M. (2011). *Scale construction and psychometrics for social and personality psychology*. London: Sage.

Gerbing, D. W., & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research*, *25*, 186-192.

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*, 930-944.

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

Goldberg, L. W., & Digman, J. M. (1994). Revealing structure in the data: Principles of exploratory factor analysis. In S. Strack & M. Lorr (Eds.), *Differentiating normal and abnormal personality* (pp. 216—242). New York: Springer.

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement, 71*, 306-324.

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

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, 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*, 319-336.

Hinz, A., Einenkel, J., Briest, S., Stolzenburg, J. U., Papsdorf, K., & Singer, S. (2012). Is it useful to calculate sum scores of the quality of life questionnaire EORTC QLQ-C30? *European Journal of Cancer Care, 21,* 677 - 683.

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*, 523-531.

Holzinger, K. J., & Swineford, F. A. (1939). *A study of factor analysis: The stability of a bi-factor solution* (No. 48). Chicago: University of Chicago Press.

Houts, C.R. & Edwards, M.C. (2019). Models for fit and models for scoring: Some thoughts on the contradictory norm within scale development.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling*, *6*, 1-55.

Hussey, I., & Hughes, S. (2019). Hidden invalidity among fifteen commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*.

Joreskog, K. G., Sorbom, D., & Magidson, J. (1979). Advances in factor analysis and structural equation models.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Westport, CT: ACE/Praeger.

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320-341.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives, 15,* 51-69.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23,* 412-433.

McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of Personality Assessment, 100,* 43-52.

Michell, J. (2012). Alfred Binet and the concept of heterogeneous orders. *Frontiers in Psychology, 3,* 1–8. https://doi.org/10.3389/fpsyg.2012.00261

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling, 2,* 255– 273.

Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences, 42,* 875-881.

Mulaik, S. (2007). There is a place for approximate fit in structural equation modelling. *Personality and Individual Differences, 42,* 883-891.

Peters, G. J. Y. (2014). The alpha and the omega of scale reliability and validity: why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *European Health Psychologist, 16,* 56-69.

Peterson, R. A., & Kim, Y. (2013). On the relationship between coefficient alpha and composite reliability. *Journal of Applied Psychology, 98,* 194-198.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12,* 287-297.

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

Rose, N., Wagner, W., Mayer, A., & Nagengast, B. (2019). Model-based manifest and latent composite scores in structural equation models. *Collabra: Psychology, 5,* 9.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48,* 1-36.

Satnor, D. A., Gregus, M., & Welch, A. (2009). Eight decades of measurement in depression. *Measurement, 4,* 135-155.

Shi, D., Lee, T., & Terry, R. A. (2018). Revisiting the model size effect in structural equation modeling. *Structural Equation Modeling, 25,* 21-40.

Sibley, C. G., Fischer, R., & Liu, J. H. (2005). Reliability and validity of the revised experiences in close relationships (ECR-R) self-report measure of adult romantic attachment. *Personality and Social Psychology Bulletin, 31,* 1524-1536.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11,* 702-712.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83,* 213-217.

Wainer, H., & Thissen, D. (1976). Three steps towards robust regression. *Psychometrika, 41,* 9-34.

Weathers, F.W., Litz, B.T., Keane, T.M., Palmieri, P.A., Marx, B.P., & Schnurr, P.P. (2013). The PTSD Checklist for *DSM-5* (PCL-5). Scale available from the National Center for PTSD at www.ptsd.va.gov.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century?. *Journal of Psychoeducational Assessment, 29,* 377-392.

Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items: Pitfalls and loopholes. *European Journal of Psychological Assessment, 31,* 231–237.

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ωh. *Applied Psychological Measurement, 30,* 121-144.