Title

Thinking About Sum Scores Yet Again, Maybe the Last Time, We Dont Know, Oh No . . .: A Comment on.

Permalink

Journal

Authors

Widaman, Keith

Revelle, William

Publication Date

DOI

Thinking About Sum Scores Yet Again, Maybe the Last Time, We Don't Know, Oh No . . .[1]: A Comment on McNeish (2023)

Keith F. Widaman[1] (iD) **and William Revelle**[2]

Abstract

The relative advantages and disadvantages of sum scores and estimated factor scores are issues of concern for substantive research in psychology. Recently, while championing estimated factor scores over sum scores, McNeish offered a trenchant rejoinder to an article by Widaman and Revelle, which had critiqued an earlier paper by McNeish and Wolf. In the recent contribution, McNeish misrepresented a number of claims by Widaman and Revelle, rendering moot his criticisms of Widaman and Revelle. Notably, McNeish chose to avoid confronting a key strength of sum scores stressed by Widaman and Revelle—the greater comparability of results across studies if sum scores are used. Instead, McNeish pivoted to present a host of simulation studies to identify relative strengths of estimated factor scores. Here, we review our prior claims and, in the process, deflect purported criticisms by McNeish. We discuss briefly issues related to simulated data and empirical data that provide evidence of strengths of each type of score. In doing so, we identified a second strength of sum scores: superior cross-validation of results across independent samples of empirical data, at least for samples of moderate size. We close with consideration of four general issues concerning sum scores and estimated factor scores that highlight the contrasts between positions offered by McNeish and by us, issues of importance when pursuing applied research in our field.

[1]University of California, Riverside, USA
[2]Northwestern University, Evanston, IL, USA

Corresponding Author:
Keith F. Widaman, School of Education, University of California, Riverside, 900 University Avenue, Riverside, CA 92521, USA.
Email: keith.widaman@ucr.edu

McNeish (2023) offered a lengthy, trenchant, and informative response to a critique that we (Widaman & Revelle, 2023) wrote of a paper by McNeish and Wolf (2020). On the surface, the McNeish (2023) riposte appears to be a tour de force contention or rebuttal of many of our arguments, supported by a series of simulations to support his points. Here, we will keep our response as short as possible, hoping to avoid having this entire discussion devolve into a tour de farce of ''they said, then we said, then he said, . . .'' ad infinitum. Such back-and-forth volleys quickly become tedious, and readers, quite reasonably, just as quickly frequently lose interest.

At the outset, we hereby state clearly that many of the claims attributed to us by McNeish (2023) were never made by us (Widaman & Revelle, 2023). Disputants often find that ''distort and distract'' is a useful rhetorical gambit: misrepresent what someone else claims and then argue against those bogus claims. We will identify a series of issues that conform to this mode of argument by McNeish (2023). In an admitted difference of opinion with McNeish, we continue to think that sum scores have sufficient psychometric properties to be useful in many research applications and may have at least one advantage over other forms of scores, such as estimated factor scores. We also always have thought that estimated factor scores might be preferable to sum scores for certain other uses, such as high-stakes decision-making (e.g., SAT scores), if the norming of estimated factor scores is based on very large samples. Thus, McNeish (2023) and we may agree on a good number of issues, even though we disagree around the edges.

Original Goal of Widaman and Revelle (2023)

Our critique (Widaman & Revelle, 2023) emphasized the issue of ''sum scores versus estimated factor scores,'' but this was due to constraints of space. Interestingly, one of our two initial intentions in preparing a critique of McNeish and Wolf (2020) was to correct a number of fundamental misrepresentations of the Holzinger and Swineford (1939) data that McNeish and Wolf used as empirical grist for their analytic mill. These issues have important implications for the empirical work and conclusions offered by McNeish and Wolf. Our concerns include, but are not limited to, the following:

1. McNeish and Wolf (2020) identified the six manifest variables used in their empirical illustrations as items and implied that they were indicators for a single latent factor. In fact, the six manifest variables were tests, each made up of many items, and they were indicators for two separate latent factors (cf. Holzinger & Swineford, 1939).

2. McNeish and Wolf (2020) implied that the six indicators had scores that fell on the same scale, stating that ''item scores range from 0 to 10.'' This is a problematic claim. Original raw scores on the six variables had very different ranges; one variable (Paragraph comprehension) had scores ranging from 0 to 19, and another (Counting dots) had scores ranging from 61 to 200. An unknown person linearly rescaled scores on each test by dividing raw scores by a different integer constant for each test so the resulting rescaled scores would fall roughly between 0 and 10; McNeish and Wolf obtained these rescaled scores through the `lavaan` package in R (Rosseel, 2012) (the raw and rescaled data are also available in the `psychTools` package in R; Revelle, 2023b). But, as examples, Paragraph comprehension had rescaled scores ranging from 0 to 6.33, and Counting dots had rescaled scores ranging from 3.05 to 10.00, so rescaled scores across the six indicators did not have strictly comparable metrics.

3. Both McNeish and Wolf (2020) and McNeish (2023) argued that scores derived from the same indicators (e.g., sum scores and estimated factor scores derived from the same items) must correlate virtually 1.00 (certainly higher than .98) to be considered to have approximately equal psychometric properties. But, McNeish and Wolf (2020) rounded the rescaled scores on the six variables from Holzinger and Swineford (1939) to integer values, such that the original scores (which had meaningful information contained in their decimal values) and the nonlinearly transformed, rounded scores on the six tests used by McNeish and Wolf tend to correlate in a range that McNeish (2023) considers unacceptable, Mdn = .967, range = .962 to .979. *Hence, readers should be advised to disregard all statistics reported by* McNeish and Wolf (2020), *because they used rounded versions of more precise scores, and the rounded scores they used had poorer psychometric properties (e.g., precision) than did the unrounded scores.*[2]

4. McNeish and Wolf (2020) advised, in general, against using sum scores due to the relatively poor psychometric properties of such scores. However, the scores on each of the six manifest variables from Holzinger and Swineford were equally weighted sums of scores on multiple items, yet McNeish and Wolf (2020) deigned to use these sum scores—albeit rescaled and then nonlinearly transformed (i.e., rounded) versions of these sum scores—as the basis of their analyses. This seems an inconsistent tack to take.

5. McNeish and Wolf (2020) misinterpreted the indicator variable for school in the Holzinger and Swineford (1939) data set. This led McNeish and Wolf to report mean differences in performance as a function of school that were in the incorrect direction, something that McNeish and Wolf might have recognized if they had consulted the Holzinger and Swineford (1939) monograph.

6. McNeish and Wolf (2020) spent some time emphasizing the rather different ability estimates for particular individuals based on sum scores across the six variables versus estimated factor scores from a one-factor congeneric model.

McNeish and Wolf preferred the model-based estimated factor scores and claimed that sum scores distorted the more precise individual differences provided by estimated factor scores, to the detriment of sum scores. This demonstration was not, however, uniquely problematic for sum scores, because McNeish and Wolf also argued at another point that a sum score across a set of indicators is perfectly correlated with estimated factor scores based on a one-factor parallel test model for those indicators. Hence, their contrast between sum scores and estimated factor scores—touted as very negative for sum scores—could be recast legitimately as the different estimated ability levels for particular persons derived from a one-factor congeneric model versus a one-factor parallel test model based on the same indicators. In addition, this entire demonstration was moot, as neither the one-factor congeneric nor the one-factor parallel test model had anything close to acceptable fit to the data. Thus, neither sum scores across the six indicators nor estimated factor scores from one-factor models fit to the Holzinger and Swineford (1939) data have any psychometric justification or valid interpretation whatsoever.

The preceding points are probably sufficient to indicate our concerns about how McNeish and Wolf (2020) misrepresented and misused data derived from Holzinger and Swineford (1939). Additional details about these matters are contained in Supplementary Material for the Widaman and Revelle (2023) paper at https://doi. org/10.3758/S13428-022-01849-w. Rather than acknowledge any of these problems, McNeish (2023) essentially moved the goal posts—pivoting away from our concerns about Holzinger and Swineford data and toward disputing our positions, supporting his claims with results from simulations. To these issues we now turn.

Misrepresentations by McNeish (2023) of Our (Widaman & Revelle, 2023) Views

In addition to correcting misrepresentations of the Holzinger and Swineford (1939) data, a second goal of the current paper is to correct misrepresentations of our positions by McNeish (2023). The second impetus for writing our original critique (Widaman & Revelle, 2023) was to argue that sum scores do not have all of the negative qualities ascribed by McNeish and Wolf (2020), especially in contrast to their unbridled preference for estimated factor scores. To be clear, we have never thought that sum scores were perfect or would be preferable to estimated factor scores in all instances. In fact, we think that sum scores and estimated factor scores each have strengths and weaknesses. We were trying simply to level the playing field with regard to these approaches, to reinforce the idea that both have their positive characteristics and both have their weaknesses, and to encourage researchers to attend carefully to the positive and negative qualities of both sum scores and estimated factor scores when pursuing their analyses.

Here, we discuss briefly a series of instances in which McNeish (2023) attributed certain views to us that we had not expressed and would not avow. A first example is the claim by McNeish (p. 2) that

> Widaman and Revelle understandably considered the models in our original article to be overly and unnecessarily restrictive because—as noted by Widaman and Revelle—the constraints within our models are not necessary to define the reliability of a sum score.

This is an incorrect statement of our position. We presume that McNeish here referred to our comments about the inapplicability of quite restrictive parallel test models that McNeish and Wolf (2020) fit to the Holzinger and Swineford (1939) data. Our opposition to the fitting of parallel test models to those data has nothing to do with whether one could define reliability of a sum score based on a less restrictive factor model. Instead, we hold that parallel test models should never be fit to the six variables from Holzinger and Swineford (1939) data because the indicators (a) were never designed to be parallel tests and (b) do not have the characteristics required of parallel tests.

As we noted (Widaman & Revelle, 2023), parallel test models assume that tests (or indicators) have equal means, equal *SD*s, equal factor loadings, and equal error variances, hence equal reliabilities. Parallel tests also typically have content specifications that ensure that comparable material is contained across the parallel tests. The six tests from Holzinger and Swineford meet none of these conditions. Raw scores on the six variables under consideration had very different raw score means and *SD*s, and Holzinger and Swineford never attempted to design the indicators to have comparable content. Instead, the tests were designed to be somewhat different, but still indicators for specifiable common latent variables or factors. For example, Holzinger and Swineford could have developed three separate tests of Paragraph Comprehension, and then parallel test constraints might be appropriate with proper content specification. But, this would lead to identification of a very narrow ability factor for Paragraph Comprehension, rather than the broader Verbal Comprehension dimension they intended to investigate. Instead of narrowly defined tests, the three indicators of Paragraph Comprehension, Sentence Completion, and Word Meaning all appear to require a common ability for extracting meaning from text, a factor identified as Verbal Comprehension, yet offered alternative and only partially overlapping approaches to assessing the common latent variable. Moreover, these tests did so with different numbers of items, markedly differing means and *SD*s, and differing reliabilities, all suggesting that an investigator would never consider parallel test constraints for this set of indicators based on raw score distributions for the variables. The haphazard, ad hoc rescaling of scores on the three indicators—conducted by an anonymous researcher and made available through the `lavaan` package—did indeed make the *SD*s on the variables more comparable. But, ad hoc rescaling of variables to essentially equalize their very unequal raw score means and *SD*s does not make them parallel tests. Furthermore, proceeding to fit models that

embody the assumption that the *SD*s are equal amounts to unacceptable capitalization on sample information.[3] To be clear, in our opinion, parallel test models are not appropriately fit to the Holzinger and Swineford data because of the nature of the variables involved, independent of any decision regarding whether reliability could be estimated from a less restrictive common factor model fit to those data.

As a second example, McNeish (2023) stated that ''Widaman and Revelle say that the ability to separate true score variance from error variance justifies sum scoring.''*We never said that.* In our opinion, scoring of a set of items must be justified on some basis, and an investigator should provide this justification in some fashion. A common justification for differential weighting of items is a pattern of differential loadings of items on a congeneric factor or the superior fit of a two-parameter item response theory (IRT) model over a one-parameter IRT model, as McNeish rightly argued. On the other hand, one might decide to use equal weighting of items if factor loadings did not differ too much on a common factor and if the goal was to ensure greater comparability of scores across independent studies that use the same items. If an investigator wishes to use an equally weighted sum of a set of items and if a one-factor model has adequate fit to the data, then the parameter estimates from the model can be used to estimate reliability of the sum score, and the resulting reliability coefficient goes by the name of coefficient ω_T. But, we reiterate that the decision on how to score a set of items should be justified on some basis, and we consider common factor analysis and IRT modeling to provide useful information in this regard. Using such information, an investigator should decide how to combine scores across a set of items and then justify that decision. Once the decision is made, an appropriate approach to estimating reliability is often extremely informative, even if the ability to estimate reliability does not itself justify any particular form of scoring.

On a third, related note, McNeish (2023) quoted Borsboom and Mellenbergh (2002) about the distinction between true scores and what they called construct scores. *Here, McNeish stated that we* (Widaman & Revelle, 2023) *conflated reliability and validity; we dispute this assertion.* We agree with McNeish that we may have used outdated and potentially problematic terms when referring to *true scores* in the context of *classical test theory* (CTT), having done so because many practicing scientists have at least passing acquaintance with such terms, and we wanted to communicate with them. Among others, Borsboom and Mellenbergh discussed rather technical, crucial, yet arcane conceptual problems with the definition and conception of true scores within CTT. They clearly preferred theoretical notions from what is known as modern test theory, which subsumes both confirmatory factor analysis (CFA) and IRT. Modern test theory approaches involve psychometric models that are specifiable and rejectable models, something not included in classic references in CTT. We hasten to add that we agree strongly with McNeish and with Borsboom and Mellenbergh on this point, and we prefer to use terms such as *latent variable scores* or *factor scores* (rather than *true scores*) when referring to the theoretical

scores on the latent factors in CFA models, and to contrast these with *manifest variable scores* that are scores on the indicators of the factors. In addition, we framed our discussion of optimal ways of estimating reliability in just this fashion, by setting up the problem as a CFA model (see our Equation 6, in Widaman & Revelle, 2023), consistent with Borsboom and Mellenbergh. Moreover, we (e.g., Revelle & Condon, 2019; Revelle & Zinbarg, 2009; Widaman & Revelle, 2023; Zinbarg et al., 2005, 2006) have long stressed the need to investigate or verify the dimensional structure of a set of items or indicators prior to deciding how to score the indicators and then estimate reliability. We admit to remaining a bit flummoxed by the apparent contention by McNeish (2023) that the scoring model (i.e., how items are to be combined) should precede the fitting of a factor model to the data. We continue to hold that the factor structure of a set of items should be evaluated first using CFA or IRT modeling, and these results should inform a researcher on how scores should be constructed from the items. Because this is a bit of a side issue and given constraints of space, we leave this latter issue for further consideration by readers.

Too little information was provided by McNeish (2023) to make much sense with any assurance of the distinction by Borsboom and Mellenbergh (2002) between true scores and construct scores, especially as this might have any relation to claims made by us (Widaman & Revelle, 2023). Borsboom and Mellenbergh argued that a true score is ''an entirely syntactic concept'' because it is ''defined in terms of the mathematical syntax of classical test theory.'' We agree and then hasten to point out that the latent variable scores of modern test theory must also be entirely syntactic concepts because they are defined in terms of the mathematical syntax of CFA and IRT models. Borsboom and Mellenbergh then argued that one should never equate the true scores (or, presumably, latent variable scores) that are mathematical placeholders in statistical models with construct scores, because the latter are semantic concepts that depend on the meaning assigned to them in relation to a construct. Once again, we agree. But, we do not understand how these distinctions relate to arguments by Widaman and Revelle (2023). Latent variable scores reside within specifications of CFA and IRT models. If a one-factor CFA model, for example, fits a data set very well, this does not mean that one has established precisely what the latent variable score represents. ''What it represents'' is a task for validity evaluation, which concerns a number of things including how the latent variable score is related to a theoretical construct and how the latent variable score, as a construct score, is related to scores for other constructs. Empirical relations among construct scores can be pursued within structural equation models (without having to estimate factor scores for individuals) or by estimating construct scores in some fashion and relating these scores to other variables to which they should, in theory, be related. The successful isolation of a latent variable and its implied or estimated scores within a CFA or IRT model should not be an end in itself, but merely sets the stage for validation work that is the heart of substantive research in psychology.

As one final note on this topic, McNeish (2023) appeared to chide us for paying attention to optimal ways to estimate reliability, rather than focusing on validity. We

certainly agree with McNeish and others (e.g., Borsboom & Mellenbergh, 2002) that validity is the core issue that should motivate empirical research. But, while emphasizing validity, one should not throw out or dismiss considerations of reliability when concentrating on sophisticated latent variable modeling. Both CFA and IRT models of modern test theory retain informative methods of estimating reliability, and appropriate indices of reliability should routinely be reported for the scores used in empirical research, be they sum scores or estimated factor scores.

Next, when contrasting the performance of sum scores versus factor scores, McNeish (2023) claimed we stated that, ''once a researcher has engaged in some psychometric work to determine the factor structure, then unweighted sum scoring is adequate,'' and referred to the last line of the Abstract in Widaman and Revelle (2023). *This is a motivated exaggeration of what we said*. What we said was, ''once the dimensional structure of a set of items is verified, sum scores *often* have a solid psychometric basis and therefore are *frequently* quite adequate for psychological research'' [emphasis added]. We do not think that sum scores are always adequate, as McNeish mistakenly implied. Indeed, in closing our paper, after wishing sophisticated latent variable methods well in the future, we wrote ''but, long live sum scores (properly vetted)!'' Hence, we think that sum scores should be vetted, even vetted quite severely (cf. Mayo, 2018), prior to their use. One should engage in psychometric work of various kinds to ensure that sum scoring is a reasonable approach for a particular set of items. One can check things like the approximate equality of factor loadings of items on a common factor, whether deleting low loading items would improve reliability and so on. Once reasonable and comprehensive vetting steps are conducted, we still feel that sum scores will frequently—not always (!)—be an adequate basis for pursuing many—certainly not all (!)—forms of empirical research.

One additional motivated exaggeration of what we said occurred when McNeish (2023) stated that ''Widaman and Revelle (2023) suggested that sum scoring is desirable because it has less sampling variability.'' McNeish left the ''it'' to which he referred unspecified. We had argued that the one form of lowered sampling variability of interest to us was the lowered sampling variability of the weights used to sum up the scores. This is demonstrably true, as unit weights used in typical sum scoring have zero sampling variability—they are all 1.0 in all samples—whereas factor scoring weights used in factor score estimation vary from sample to sample. We think it is fine that McNeish (2023) investigated additional forms of sampling variability, but these latter results have essentially nothing to do with regard to the claims we made.

Loose Terminology Sinks Precision

Just as loose lips are said to sink ships, loose terminology sinks precision. Certain terminology used by McNeish (2023) seems at least potentially problematic, so deserves mention. First, recall that McNeish kept emphasizing that we inadvisably discussed *true scores* and how one might use estimated effects of true scores to separate true score variance from error variance, and we mentioned CTT in passing. McNeish

chided us because *true score* is a problematic concept in the context of CTT, citing Borsboom and Mellenbergh (2002). However, later, when discussing his simulation procedures, McNeish stated that one advantage of simulation methods is that ''unlike real empirical data, it allows the truth to be known.'' A bit later, he described how, in his data simulation process, ''true scores were sampled from a standard normal distribution.'' Still later, McNeish opined that ''the benefit of a simulation study is that the true scores are known.'' In a footnote, McNeish stated that he was using the term *true score* to refer to the correct score when generating data in a simulation model, not to a concept in CTT. This serves to underscore our presumption, stated above, that *true scores* are apparently fine and dandy if considered within the mathematical formulations of CFA or IRT models of modern test theory, but are outmoded and problematic notions in the context of CTT. This is certainly a relief to us, as we developed our discussion of reliability estimation within the context of CFA models (see Equation 6 of Widaman & Revelle, 2023). Thus, despite our potentially ill-advised nod toward CTT, our use of the term *true score* (or, more precisely, *true factor score*) in the CFA context appears to be eminently acceptable. Moreover, the ability to separate true score variance from error variance is still possible and very informative within both CFA and IRT models of modern test theory, so highlighting issues of reliability and associated measurement precision is not to be scoffed at.

Next, McNeish (2023) failed to qualify adequately his use of the term *factor scores* in most places in the manuscript. For example, in the process of simulating data, he stated that ''factor scores were computed with the regression approach.'' An accurate wording would be that factor scores were *estimated* (i.e., not *computed*) with the regression approach. Several different methods can be used to estimate factor scores, the regression method is one of these, yet the different methods tend to yield somewhat different estimates of the factor scores. Furthermore, when discussing the derived scores in simulations, one should be careful to continue referring to them as *estimated* factor scores. Requiring use of the term *estimated* every time one refers to estimated factor scores can perhaps seem tedious and stilted, but reinforces the notion that these are estimated factor scores, not true scores in any sense. Moreover, use of a different method of factor score estimation might lead to somewhat different results. On occasion, McNeish (2023) used the wording of *estimated factor scores*, but this was the exception, not the rule.

Data Analyses Should Be Based on Reality

Simulated Data

We turn next to the issue of simulations of data, a laudable approach in general. Simulations are a powerful tool for examining the utility or validity of quantitative techniques. In his response (McNeish, 2023) to our prior critique (Widaman & Revelle, 2023), McNeish appeared to show that, even if estimated factor scores and unit-weighted sum scores correlate .96, estimated factor scores still achieve better fit to the underlying true factor scores. To be convincing, simulations should be based

on plausible parameter values and simulation methods. Unfortunately, not all simulations by McNeish adhered to this proviso.

Due to the limitations of space, here we can make only a few, less-than-detailed comments on the McNeish (2023) simulations. We provide more complete, detailed critique in Supplemental Material, available at https://osf.io/ayn2m/. Two major problems are present in the McNeish simulations. First, and importantly, McNeish simulated data with perfect fit in the population (i.e., with no forms of model misfit in the population). Empirical data appear to contain misfit in the population (cf. Tucker et al., 1969), so simulations are most informative if they contrast outcomes under different levels of model misfit in the population.

Second, certain conditions in the McNeish (2023) simulations are not representative at all of empirical research situations. The most problematic condition was the one with 15 manifest indicators designed so factor scores and sum scores would correlate only .96 in the population. In this condition, seven of the indicators had standardized loadings of .95, one had a loading of .40, and seven had loadings of .30. One way of vetting sum scores is to estimate reliability and to see whether deleting certain items would increase reliability. In the condition under consideration, coefficient ω_T for the 15 indicators based on the population loadings is .914, a surprisingly low value given the very high loadings for seven of the items. Simple calculations indicate that coefficient ω_T would improve substantially to .985 if one deleted the eight items with low loadings and retained only the seven items with the high loadings. This condition was the one in which sum scores fared the worst in comparison to estimated factor scores, yet proper and severe vetting of sum scores would lead to a rejection of sum scores across all 15 indicators.

We feel no need to ''recreate a wheel'' in simulating data if good alternatives already exist. Here, we refer to a very recent paper by Liu and Pek (2023), who designed a series of simulations that were based on empirical research and prior simulation studies. Liu and Pek systematically varied levels of measurement error (the three levels of high, low, and varying indicator factor loadings), sampling error (i.e., three different sample sizes), and model error (five levels, from no model error to multiple crossed loadings unmodeled). In general, summed scores and estimated factor scores performed in a similar fashion. Summed scores appeared to be more robust to measurement error and model error in many conditions. On the other hand, estimated factor scores performed better in certain conditions, but only when model error was small and sample size was large. Liu and Pek found no clear winner overall, but different strengths and weaknesses of summed scores and estimated factor scores.

The upshot of the comparison of simulations by McNeish (2023) and Liu and Pek (2023) is the following: McNeish employed conditions that were not generally representative of empirical research or prior simulations, but were formulated to lead to differences in performance by estimated factor scores and sum scores, which is precisely what he found. In contrast, Liu and Pek used conditions that were more generally representative of empirical research and prior simulations, employed informative

and varying forms of model misfit, and found notable strengths and weaknesses of both sum scores and estimated factor scores in different conditions. Caveat emptor with regard to results of simulations!

Empirical Data

To supplement simulated data results, we conducted analyses of empirical data to compare estimated factor scores and sum scores, using the 4,000 cases in the `spi` data set[4] from the `psychTools` package (Revelle, 2023b) in R. We used the 70 items for assessing the Big Five factors of personality—Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. Using methods from the `psych` package (Revelle, 2023a) in R, we obtained a five-factor solution for the 70 personality items and estimated factor scores with the regression method. Correlations between estimated factor scores and equally weighted sum scores were uniformly high, $r(3998) = .97$ for each of the five dimensions.

Of greater interest, we investigated cross-validation of results when using estimated factor scores or sum scores to predict criteria. The `spi` data set includes eight criteria, scored in the following fashion:

(a) parent 1 education, from 1 = *less than 12 years* to 8 = *graduate or professional degree*;
(b) parent 2 education, same scale as parent 1 education (a);
(c) emergency room visits in the past year, from 1 = *none* to 4 = *3 or more times*;
(d) smoke, from 1 = *never* to 9 = *over 20 times per day*;
(e) own education, same scale as parent 1 education (a);
(f) exercise, scale from 1 = *very rarely* to 6 = *more than 5 times per week*;
(g) age, in years;
(h) health, on a scale from 1 = *poor* to 5 = *excellent*.

We randomly selected 250 participants from the `spi` data set to serve as the derivation sample, obtained a five-factor solution for the 70 personality items, and estimated factor scores for the factors using the regression method. We also computed five equally weighted sum scores, one for each dimension. We separately used the set of five estimated factor scores and the set of five sum scores as predictors of the eight criteria. Then, we employed data from the remaining 3,750 participants as a cross-validation sample, using the factor weights and regression weights from the derivation sample to cross-validate the regression equations separately for estimated factor scores and sum scores. We repeated this process 100 times to obtain mean multiple correlations and the *SD* of multiple correlations for both the derivation and cross-validation samples across the 100 replications. Then, we repeated this entire approach with derivation samples of

size 1,000 participants, so cross-validation sample sizes of 3,000, again repeating the random selection of derivation sample participants across 100 replications.

Results of these analyses are shown in Table 1 and Figure 1. In the top half of Table 1, results for derivation samples with 250 participants are shown. For derivation sample analyses, estimated factor scores led to slightly higher multiple correlations when predicting criteria relative to the performance by sum scores, although differences were very small. But, the cross-validation results offer a clearer contrast, as the equally weighted sum scores had higher cross-validated multiple correlations for all eight criteria, differences that were largest for criteria with the largest multiple correlations. The bottom half of Table 1 contains results for the larger derivation samples having 1,000 participants. Here, once again, estimated factors scores had slightly higher multiple correlations when predicting criteria compared with the equally weighted sum scores in the derivation sample. But, again, the cross-validation results show the opposite trend, with equally weighted sum scores performing as well as or slightly better than estimated factor scores.

The results of the cross-validation results are shown more strikingly in Figure 1, with criteria arrayed from left to right as a function of magnitude of the cross-

Table 1. Multiple Correlations Predicting Eight Criteria From Estimated Factor Scores or Sum Scores, For Derivation Sample Sizes of 250 and 1,000.

Sample size[a]	Criterion[b]	Derivation sample		Cross-validation sample	
		Factor[c]	Sum[d]	Factor[c]	Sum[d]
250	p1edu	.18 (.05)	.18 (.05)	.04 (.03)	.05 (.03)
	p2edu	.19 (.05)	.19 (.06)	.05 (.03)	.05 (.03)
	ER	.21 (.06)	.20 (.06)	.08 (.03)	.08 (.03)
	smoke	.23 (.06)	.23 (.06)	.13 (.03)	13 (.03)
	education	.30 (.06)	.29 (.06)	.21 (.03)	**.23** (.02)
	exercise	.30 (.06)	.29 (.06)	.21 (.03)	**.23** (.03)
	age	.33 (.05)	.32 (.05)	.26 (.03)	**.28** (.02)
	health	.43 (.05)	.43 (.05)	.34 (.02)	**.39** (.01)
1,000	p1edu	.12 (.03)	.12 (.03)	.07 (.02)	.07 (.01)
	p2edu	.13 (.02)	.13 (.02)	.08 (.01)	.08 (.01)
	ER	.15 (.03)	.14 (.03)	.12 (.01)	.11 (.01)
	smoke	.19 (.03)	.19 (.03)	.17 (.01)	.17 (.01)
	education	.27 (.03)	.27 (.03)	.25 (.01)	.25 (.01)
	exercise	.28 (.03)	.28 (.03)	.25 (.01)	.26 (.01)
	age	**.32** (.03)	.31 (.03)	.30 (.02)	.30 (.01)
	health	.41 (.02)	.41 (.02)	.38 (.01)	**.40** (.01)

Note. Tabled values are mean multiple correlations, with SDs in parentheses. Bold-faced coefficients represent mean differences of .01 or greater between estimated factor scores and sum scores. [a] Sample size is size of derivation sample. [b] Criteria (see text for coding of variables): p1edu = parent 1 education; p2edu = parent 2 education; ER = emergency room visits. [c] Factor = estimated factor score. [d] Sum = equally weighted sum score.
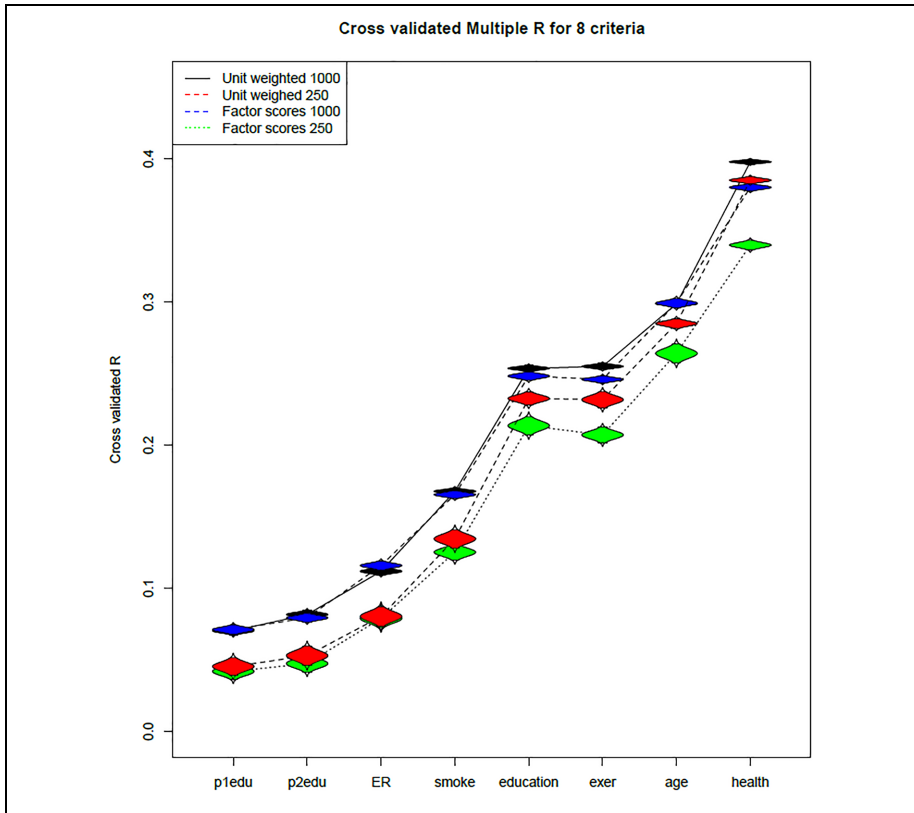
Figure 1. Resampled Cross-Validated Multiple Correlations for Predicting Eight Criteria From Estimated Factor Scores and From Sum Scores.

Note. [Green = estimated factor scores, $N = 250$; Red = equally weighted sum scores, $N = 250$; Blue = estimated factor scores, $N = 1,000$; Black = equally weighted sum scores, $N = 1,000$] [Criteria: p1edu = parent 1 education; p2edu = parent 2 education; ER = emergency room visits; exer = exercise].

validated multiple R. For the $N = 250$ conditions, the average multiple Rs for unit-weighted sum scores (in red with dashed lines) always exceeded those for estimated factor scores (in green with dotted lines), with the relative advantage of sum scores increasing as the multiple R for a criterion increases. For the $N = 1,000$ conditions, results were not as marked, but were similar in general pattern. For the first four criteria (on the left half of Figure 1), unit-weighted sum scores (in black, solid lines) had cross-validated multiple Rs of similar magnitude to those for estimated factor scores (in blue, dashed lines). But, for the last four criteria (right half of Figure 1), sum scores tended to perform as well as or better than estimated factor scores, with the difference increasing as a function of magnitude of the multiple R.

At least two conclusions are supported by our analyses of empirical data. One conclusion is that estimated factor scores may outperform sum scores in the sample on which the factor scoring coefficients and resulting factor scores have been estimated, but may lead to poorer cross-validation of results in new samples relative to performance by sum scores (cf. Table 1). This relative disadvantage of estimated factor scores in the cross-validation context is likely to diminish as derivation sample sizes increase (cf. Figure 1). At present, we are not aware of how large derivation samples must be to ensure more adequate cross-validity for estimated factor scores; this clearly is an issue that deserves additional work in the future.

The second conclusion is that analysis of empirical data is a useful, perhaps necessary adjunct to the analysis of simulated data. Simulations can offer insight regarding perfect worlds in which all assumptions of methods are met, whereas empirical data analyses bring us back down to earth to deal with real data, warts and all. If analyses of simulated data and analyses of empirical data arrive at similar conclusions, this welcome result bolsters faith in both types of data. However, if simulated and empirical data analyses lead to different outcomes, researchers should be warned not to take simulated data results too much to heart, as the important validation work we pursue in applied, substantive research will be based always on empirical data.

The Crux (or Cruxes) of the Matter

In this section, we concentrate on major issues that offer a contrast between McNeish (2023) and us. We felt it unwise to attempt to respond to each and every point made by McNeish, as this would require altogether too much time and journal space. Instead, in light of the foregoing, we highlight here four major points worth considering.

Issue 1: Alignment of Factor and Scoring Model

McNeish (2023) stated that a ''fundamental source of disagreement [between McNeish and Wolf (2020) and Widaman and Revelle (2023)] seems to be whether the factor model (whose fit is being assessed) must align with the scoring model when justifying sum scores.'' This was and remains an issue of some disagreement, although not total disagreement. Much can be said in support of having a factor model and scoring model align, and McNeish did a fine job of outlining arguments in favor of alignment. We acknowledge that alignment of factor model and scoring model is preferred in certain, perhaps many, situations. Thus, if loadings on a latent factor differ, alignment of the scoring model would dictate that unequal weights are used when estimating factor scores. This might be the best way to proceed in a given research application. McNeish cited a paper by Thissen et al. (1983) that argued in favor of alignment of factor and scoring models. Specifically, Thissen et al. argued that equal weights could be used if a one-factor factor model with equal loadings fit data adequately, but unequal weights were preferred if the one-factor model had

unequal loadings. Importantly, Thissen et al. added that the unequal weights would be based on sample data and these weights were unlikely to generalize across samples, an issue of particular importance to us. If scoring weights do not generalize across samples, comparisons across samples will be compromised, perhaps in major ways.

The crux of our disagreement on this issue is whether one must *always* align the factor model and scoring model in *all* research applications. McNeish (2023) appears to think that such alignment must always occur; we think that alignment is not always necessary, although one should be aware of any potential problems associated with the use of sum scores computed using equal weights. As a hypothetical example, suppose that one had 18 manifest variables—three indicators for each of six latent variables—with a restricted pattern of relations hypothesized to hold among the six latent variables. Consider Alternative 1: If each indicator were a scale composed of multiple items (e.g., 12 items), the McNeish (2023) position, apparently, is that one should conduct 18 preliminary sets of analyses, one for each indicator. Each preliminary set of analyses would involve fitting various models to the items for a given indicator, including perhaps one-factor parallel, tau equivalent, and congeneric test models and/or one-parameter and two-parameter IRT models. If one-factor models did not fit adequately, then models with more than one factor should be evaluated. After all of this work is done, one would then select the optimal model for items for the particular indicator, design a scoring model aligned with the latent variable model, and estimate factor scores. After this voluminous work was done for each of the 18 indicators, one would use the 18 estimated factor scores as indicators in the structural equation model motivating the research. Now, consider Alternative 2: if the 18 indicators had been developed in prior measurement work and found to be essentially unidimensional, a researcher could simply form 18 sum scores, one equally weighted sum score for each indicator, and get on with fitting the structural equation model. In apparent disagreement with McNeish, in many research situations we think Alternative 2 is a reasonable one to take, even if not perfect in all respects.

As a quick aside, assume that an investigator still thinks Alternative 1 is the preferred approach. Suppose the researcher has both males and females in the sample and intends to investigate sex differences in the restricted relations among the six latent variables. Would this investigator feel compelled to engage in the 18 sets of preliminary analyses—one for each indicator—separately for males and females prior to estimating factors scores that would enable the two-group structural modeling to commence? And, if they computed best-fit parameter estimates and thus different scoring coefficients separately for each indicator in each of the two groups, how would this researcher ensure that estimated factor scores were on comparable scales across groups, which would be required to make interpretable comparisons across groups on estimates of structural model parameters? This researcher might be envious of a second investigator who preferred Alternative 2, so computed equally weighted sum scores on the 18 indicators for all participants and moved directly to the two-group structural modeling. The second investigator used sum scores that,

admittedly, rest on a number of untested assumptions (e.g., measurement invariance for each indicator across groups). If any assumptions were invalid, the results using sum scores might be compromised. In an era of open science transparency, interested researchers could follow up the analysis by the second investigator to determine whether implicit assumptions were reasonable or whether the analysis was fatally flawed.

Issue 2: Comparability of Scores Across Studies

A second, fundamental issue is the issue of comparability of scores across studies, given our contention that sum scores yield more comparable results across studies. We highlighted this issue as perhaps the most salient advantage of sum scores over estimated factor scores. To quote Widaman and Revelle (2023), ''if sum scores are composed in exactly the same fashion across studies, the results of the studies will be more easily compared than if different weighting were used across studies.'' This claim is based on a number of assumptions or requirements, such as, across studies, the same number of items is used, the wording of items is the same, and the response scale is also identical. As pointed out by McNeish (2023), the use of sum scores also assumes that measurement invariance holds across samples, although we think that approximate invariance may be sufficient. Admittedly, the comparability of sum scores across studies is at a manifest variable level, because sum scores are manifest variable scores formed as equally weighted sums of items or indicators. Note that estimated factor scores are also, in essence, manifest variable scores, but estimated as differentially weighted sums of sample-mean centered and standardized item or indicator scores. We contend that use of equal weights allows easier comparisons across samples or studies and that differential weights across samples or studies makes comparisons difficult or impossible. *We could not find that McNeish (2023) ever offered clear rebuttal or even consideration of this point, a central claim in our Widaman and Revelle (2023) paper.*

McNeish (2023) made a number of additional, very cogent observations about problems with sum scores and thus comparative strengths of estimated factor scores, and we agree with many of them. For example, McNeish offered very clear discussion about how the use of sum scores assumes measurement invariance across groups. If sum scores are used in situations in which measurement invariance fails in important ways, comparative results across samples will be compromised. This concern regarding measurement invariance is particularly important when the groups are from different cultural groups or strata that may respond to wording in items in different ways. In such situations, use of sum scores may indeed be problematic. In discussing this point, McNeish referred to the ''crying'' item from the Beck Depression Inventory-2 (BDI-2). Males and females apparently respond to this item rather differently, and this form of differential item functioning (DIF) could have an effect on male-female comparisons on a sum score across the 21 items of the BDI-2. We think the proof is in the pudding on matters of this sort. Some authors have found that DIF

sometimes cancels out when estimating a scale score (e.g., Stark et al., 2004), but one certainly cannot count on this happening. Thus, we think that results could, and perhaps should, be compared when estimated latent variable scores are used that embody item DIF (if such is present) and when simple sum scores are used in analyses. If results differ in notable ways, arguments about the inappropriateness of using sum scores gain strength. If results differ little or essentially not at all, arguments against sum scores lose their force. Again, proof is in the pudding, not in a priori declaration by fiat.

Issue 3: Maximizing Model Fit in a Sample Versus Promoting Replication Across Studies

We think that a third, very important, fundamental difference of opinion between McNeish (2023; McNeish & Wolf, 2020) and us stems from the issue of maximizing fit to data in an individual sample versus providing a basis for replicating results across samples. Whenever one fits any statistical model to data, one should be aware that parameter estimates and model fit are optimized as the model is fit to data. We typically assume that a statistical model we use is primarily fitting real empirical trend in the data, but should always be aware that optimal fit to sample data also implicitly involves fitting some error in the sample, error that would not reoccur in another sample. Recent work on fungible regression weights by Waller (2008) demonstrates that one should not trust ''best fit'' sample estimates too much, as alternate sets of regression weights—many of which do not closely resemble the best-fit estimates—provide almost as good fit to the data as do the best-fit estimates. McNeish (2023) used best-fit sample estimates to make his points, and we think that replication and/or cross-validation of results should be given at least equal weight to the issue of optimal fit to data in a single sample.

Our results offer clear ammunition on this point, based on our analyses of empirical data. With a moderate sample size of 250 participants in a derivation sample, we found that estimated factor scores used as predictors tended to have slightly higher prediction of criteria than did sum scores when used as predictors in the derivation sample, although differences were not large. But, in cross-validation analyses, sum scores outperformed estimated factor scores, yielding higher cross-validated multiple correlations for all eight criteria, especially for criteria with larger multiple correlations. With the larger derivation sample size of 1,000 participants, differences between estimated factor scores and sum scores were attenuated, but sum scores still performed as well as or better than estimated factor scores in cross-validation analyses.

The most reasonable conclusion on this issue is that the size of the derivation sample and the nature of the items are probably both very important. If derivation sample sizes are very large, as in many state or national educational testing situations, and if items are carefully constructed, then estimated latent variable scores may well perform as well as or better than simple sum scores in additional samples. But, with

sample sizes in the range encountered in much substantive work in psychology, with sample sizes ranging at most up to 1,000 participants, little in the way of clear preference for estimated latent variable scores over simple sum scores can be found. Indeed, sum scores outperformed estimated factor scores in cross-validation when derivation sample sizes were moderately large (i.e., 250–1,000 participants), which should be an outcome of some concern for proponents of estimated factor scores.

Issue 4: Sample Standardization and Relations With External Variables

A fourth and final matter is the sample-based standardization that occurs when estimating factor scores. We think it is possible to *lose oneself in explanatory space* if one always standardizes manifest variable scores to $M = 0.0$ and $SD = 1.00$ in a sample, as is typically done when estimating factor scores. What we mean by *lose oneself in explanatory space* is that centering data always to the centroid of sample space makes both characterization of the current sample and comparisons with other samples or the population difficult or impossible. Samples differ in mean and variance on items and other manifest variable raw scores, at times differing substantially. Standardizing variables to $M = 0.0$ and $SD = 1.00$ within samples essentially discards these differences.

Return to consideration of the well-known depression instrument, the BDI-2. The BDI-2 contains 21 items, each scored on a 0 to 3 scale, so raw sum scores can vary between 0 and 63. Many articles have reported factor analyses of BDI-2 items (e.g., Brouwer et al., 2013; Osman et al., 1997; Subica et al., 2014; Ward, 2006). The general finding is that, when oblique first-order factors were estimated for BDI-2 items, correlations among factors tend to be very high (e.g., frequently above .80). When bifactor models were fit, the general factor predominated, with little usable variance left over on orthogonal group factors. As a result, most authors recommended that, because the general factor predominates, the overall score across the 21 items seemed the most reasonable and interpretable score to use.

But, the samples differed considerably in mean level based on raw sum scores. Consider the following two examples, which are not the most extreme to be found in the literature. The Osman et al. (1997) sample of 230 college students had a pooled mean (across males and females) of 11.09 and pooled within-group $SD = 8.01$. In comparison, the sample of 1,904 patients used by Subica et al. (2014) had much higher scores, $M = 24.93$, with rather larger variance, $SD = 13.00$. The BDI-2 manual outlines score ranges that indicate severity of depression problems, with raw scores of 0 to 13 indicating minimal depression, 14 to 19 mild depression, 20 to 28 moderate depression, and 29 to 63 severe depression. The mean score in the Osman et al. study was clearly in the minimal range of depression, and, given the SD, one would expect few participants to score in the severe range. In contrast, the mean score in the Subica et al. study was above the midpoint of the moderate depression range and one would expect a substantial proportion, perhaps 20% to 30% or even more, to fall in the severe depression range. Estimating factor scores in the typical fashion would

lead to scores with $M = 0.0$ and $SD = 1.00$ (approximately) within each sample, erasing the implications explicit in the raw scores. The top 5% (or other percent) in each sample would still be the subsample with the highest depression scores within the given sample. But, the top 5% of the Osman et al. sample would have very different likelihood of severe depression problems than would the top 5% of the Subica et al. sample.

As a final consideration here, McNeish (2023) went to great lengths to identify persons with true (known) factor scores at a given extreme value (e.g., the top 5% of the sample) and then compared estimated factor scores and sum scores with regard to their sensitivity and specificity with regard to correctly characterizing individuals who did or did not fall in the clinically significant range. In these analyses, sum scores had noticeably worse performance than did estimated factor scores. Although this may seem a telling problem for sum scores, we hasten to add that most screening instruments currently in use employ cutoff scores based on raw scores, usually sum scores, rather than estimated factor scores. Harking back to the BDI-2 example just above, the identification of classes of persons with differential severity of depression is made on the basis of raw score ranges, not on post hoc estimated factor scores based on sample-based standardized item scores. In effect, sensitivity and specificity of estimated factor scores are largely impossible to evaluate or trust if one employs sample-mean centered indicators in the estimation of factor scores that discard information about the absolute level of responses in the sample.

Conclusion

Clearly, much remains to be done with regard to measurement issues in psychology. To avoid weaknesses of both estimated factor scores and sum scores, a number of avenues could be taken, and McNeish (2023) offered several. In our opinion, all of the approaches discussed by McNeish should be evaluated and tested severely, and additional methods might well be developed and have great promise. Indeed, we encourage others to think of additional ways to improve our measurement approaches in our field to develop optimal ways to arrive at scores from measuring instruments.

We also encourage discussions of issues that deal directly and dispassionately with key issues in measurement, rather than providing contributions that yield more heat than light (as the saying goes). As should be clear from preceding sections of this paper, McNeish (2023) misrepresented things that we (Widaman & Revelle, 2023) wrote and then argued against those (misrepresented) positions. Regardless of whether one agrees or disagrees with many arguments offered by McNeish, his arguments are largely moot with regard to our discussion in Widaman and Revelle, because the positions attributed to us by McNeish were never made by us.

To hark back to an issue discussed by Widaman and Revelle (2023) and earlier in this paper, *sum scores have one signal strength over estimated factor scores*: given reasonable caveats (e.g., same number and wording of items, same response scales), *results will be more easily compared across studies if sum scores are used than if*

estimated factor scores are used. McNeish (2023) never offered any cogent counter-arguments on this point. Our results in this paper identify a *second strength of sum scores, as sum scores tend to lead to better cross-validation of results than do estimated factor scores with small to moderate sized samples*. The field of psychology continues to be roiled by issues related to lack of replication of results across studies, and replication of results is one important aspect of validity. We should promote methods that allow us to determine whether results replicate across studies. In our opinion, sum scores can often (not always!) be an appropriate and adequate basis for determining whether results replicate across studies, across samples, and so on.

Thus, until more adequate methods for deriving scores from measurement instruments are developed and easily implemented, we reiterate our charge from Widaman and Revelle (2023): Long live sophisticated latent variable methods! But, long live sum scores (properly vetted)!

Author Contributions

Both authors collaborated on all aspects of the manuscript. The first author took primary responsibility for writing the first draft; the second author took primary responsibility in data simulations and analyses. Both authors formulated all aspects of the manuscript, edited subsequent drafts, and approved the final version for submission.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Preregistration

No preregistration was required for this paper as it did not involve the conduction of new empirical studies, but relied on archival data.

Data, Materials, and Online Resources

All raw data analyzed in the manuscript are available through the psychTools package in the R computing environment. R script files require access to the psych and psychTools packages in R. Supplemental information and the R script files that show how to access data and then run all analyses are provided on an Open Science Framework page, located at https://osf.io/ayn2m/.

Reporting

This study involved publicly available archival data; no new data were collected.

Ethical approval

No approval was required because all data are publicly available and are de-identified.

ORCID iD

Keith F. Widaman (iD) https://orcid.org/0000-0002-6424-3998

Notes

1. Adapted from lyrics to *The Last Time*, by the Rolling Stones. Available from https://www. azlyrics.com/lyrics/rollingstones/thelasttime.html. *Behavior Research Methods* declined to publish the present comment, despite the fact that the previous papers in this series (McNeish, 2023; McNeish & Wolf, 2020; Widaman & Revelle, 2023) had been published in that journal. We thank the current journal for allowing us to correct misrepresentations by McNeish (2023) of our views in Widaman and Revelle (2023) and thereby clarify important distinctions in the choice between sum scores versus estimated factor scores.

2. As one example, consider Test 09, Word Meaning. This test consisted of 50 items, each scored 0 = *incorrect*, 1 = *correct*. The original sum score from Holzinger and Swineford (1939) was approximately normally distributed, with $M = 15.30$, $SD = 7.67$, skew = +0.86, kurtosis = +0.82, and range from 1 to 43. Across the 301 participants, at least one person had a score at each value between 1 and 36, with one person each at scores of 38, 39, 41, and 43. Thus, in the original scoring, 40 different integer values indexing individual differences in performance were contained in this distribution. An unknown person then divided raw scores on Word Meaning by 7, leading to the rescaled score available through the `lavaan` package in R. This rescaled score still contained 40 different scores across the 301 participants, and it correlates 1.00 with the original raw score, so no important psychometric properties were lost in the rescaling. McNeish and Wolf (2020) then rounded these rescaled scores to integer values, so only the seven integers between 0 and 6 (inclusive) were used in their analyses. This rounded score correlated only .967 with the original raw score in Holzinger and Swineford and with the rescaled score available through `lavaan`. As one further point, McNeish and Wolf (2020) stated that some of the rescaled scores contained decimal values. In fact, all six of the rescaled scores had decimal values, as expected when an integer-valued sum score is divided by an integer greater than 1.

3. We note that, of course, McNeish and Wolf (2020) did not perform the rescalings of the six variables from Holzinger and Swineford (1939). The rescaling was done by some unnamed researcher, and the resulting scores are available through the `lavaan` package in R. That said, McNeish and Wolf did take advantage of these rescaled scores to formulate models, such as parallel test models, that we consider inappropriate when applied to these data.

4. The `spi` data set, developed by Condon (2017), contains responses by 4,000 persons to 135 personality items and also includes information on 10 external variables. Details about the `spi` data set can be obtained using the help function in R, after

installing the `psychTools` package (Revelle, 2023b). A data dictionary with the phrasing of each item is also easily accessed through `psychTools`.

References

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence*, *30*(6), 505–514. https://doi.org/10.1016/S0160-2896(02)00082-X

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). On the factor structure of the Beck Depression Inventory—II: G is the key. *Psychological Assessment*, *25*(1), 136–145. https://doi.org/10.1037/a0029228

Condon, D. M. (2017). *The SAPA personality inventory: An empirically-derived—Hierarchically-organized self-report Personality Assessment Model*. https://psyarxiv.com/sc4p9/

Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution* (Supplementary educational monographs, no. 48). Department of Education, University of Chicago. https://www.proquest.com/scholarly-journals/study-factor-analysis-stability-bi-solution/docview/615086110/se-2?accountid=14521

Liu, Y., & Pek, J. (2023). *Summed versus factor scores: Considering uncertainties when using scores* [Manuscript submitted for publication]. Department of Human Development and Quantitative Methodology, University of Maryland.

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press. https://doi.org/10.1017/9781107286184

McNeish, D. (2023). Psychometric properties of sum scores and factor scores differ even when their correlation is 0.98: A response to Widaman and Revelle. *Behavior Research Methods*, Advance online publication. https://doi.org/10.3758/s13428-022-02016-x

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Osman, A., Downs, W. R., Barrios, F. X., Kopper, B. A., Gutierrez, P. M., & Chiros, C. E. (1997). Factor structure and psychometric characteristics of the Beck Depression Inventory-II. *Journal of Psychopathology and Behavioral Assessment*, *19*(4), 359–376. https://doi.org/10.1007/BF02229026

Revelle, W. (2023a). *Psych: Procedures for personality and psychological research* (Version 2.3.3) [Computer software]. https://CRAN.R-project.org/package=psych

Revelle, W. (2023b). *PsychTools: Tools to accompany the psych package for psychological research* (Version 2.3.3) [Computer software]. https://CRAN.R-project.org/package=psychTools

Revelle, W., & Condon, D. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. Journal of Statistical Software, *48*(2), 1–36. https://www.jstatsoft.org/v48/i02/

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important. *Journal of Applied Psychology*, *89*(3), 497–508. https://doi.org/10.1037/0021-9010.89.3.497

Subica, A. M., Fowler, J. C., Elhai, J. D., Fruch, B. C., Sharp, C., Kelly, E. L., & Allen, J. G. (2014). Factor structure and diagnostic validity of the Beck Depression Inventory—II with adult clinical inpatients: Comparison to a gold-standard diagnostic interview. *Psychological Assessment*, *26*(4), 1106–1115. https://doi.org/10.1037/a0036998

Thissen, D., Steinberg, L., Pyszczynski, T., & Greenberg, J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. *Applied Psychological Measurement*, *7*(2), 211–226. https://doi.org/10.1177/014662168300700209

Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, *34*(4), 421–459. https://doi.org/10.1007/BF02290601

Waller, N. G. (2008). Fungible weights in multiple regression. *Psychometrika*, *73*(4), 691–703. https://doi.org/10.1007/s11336-008-9066-z

Ward, L. C. (2006). Comparison of factor structure models for the Beck Depression Inventory—II. *Psychological Assessment*, *18*(1), 81–88. https://doi.org/10.1037/1040-3590.18.1.81

Widaman, K. F., & Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, *55*(6), 788–806. https://doi.org/10.3758/s13428-022-01849-w

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 1–11. https://doi.org/10.1007/s11336-003-0974-7

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: Comparison of estimators for ω_H. *Applied Psychological Measurement*, *30*(2), 121–144. https://doi.org/10.1177/0146621605278814