PSYCHOMETRIKA—VOL. 89, NO. 4, 1148–1169
DECEMBER 2024
https://doi.org/10.1007/s11336-024-09988-z

PRACTICAL IMPLICATIONS OF SUM SCORES BEING PSYCHOMETRICS' GREATEST ACCOMPLISHMENT

DANIEL MCNEISH

ARIZONA STATE UNIVERSITY

This paper reflects on some practical implications of the excellent treatment of sum scoring and classical test theory (CTT) by Sijtsma et al. (Psychometrika 89(1):84–117, 2024). I have no major disagreements about the content they present and found it to be an informative clarification of the properties and possible extensions of CTT. In this paper, I focus on whether sum scores—despite their mathematical justification—are positioned to improve psychometric practice in empirical studies in psychology, education, and adjacent areas. First, I summarize recent reviews of psychometric practice in empirical studies, subsequent calls for greater psychometric transparency and validity, and how sum scores may or may not be positioned to adhere to such calls. Second, I consider limitations of sum scores for prediction, especially in the presence of common features like ordinal or Likert response scales, multidimensional constructs, and moderated or heterogeneous associations. Third, I review previous research outlining potential limitations of using sum scores as outcomes in subsequent analyses where rank ordering is not always sufficient to successfully characterize group differences or change over time. Fourth, I cover potential challenges for providing validity evidence for whether sum scores represent a single construct, particularly if one wishes to maintain minimal CTT assumptions. I conclude with thoughts about whether sum scores—even if mathematically justified—are positioned to improve psychometric practice in empirical studies.

Key words: sum scores, total score, composite score, reporting practices, CTT, classical test theory.

This paper is inspired by the recent article by Sijtsma et al. (2024), which I thoroughly enjoyed reading. I thought the paper was thought-provoking and provided a highly readable and extensive review of the issues around scoring and classical test theory (CTT) with a clarity that has eluded many other papers on the same topic. I think the paper is well-positioned to bridge some of the historical developments in psychometrics to the new generation of psychometricians and quantitative psychologists (like me) who were trained in an environment where latent variable models were well-established and the primary emphasis. Frankly, I wish that a paper like this that lucidly and succinctly distills so many topics and so much history had existed when I first learned about psychometrics.

The treatment by Sijtsma et al. (2024) solidifies the mathematical basis for CTT-based sum scores and highlights some pertinent questions that remain in the domain of CTT. Even as someone who is generally skeptical of sum scoring, I have no major issues with the content of their article. Nonetheless, given the practical focus of their article, it seems worthwhile to go beyond what mathematics might permit and to consider potential practical implications of viewing sum scoring as psychometrics' greatest accomplishment.

As a preface, I cite some of the original authors' previous work throughout the text. This is in no way intended to be a "gotcha" or an attempt to catch the authors in contradictions, especially because sentiments may have changed in the years since the cited works were published. Rather, the authors are clear communicators who have articulated certain points far clearer than I can myself, so I rely on the original phrasing to best convey certain ideas. Additionally, the authors' previous work has greatly influenced the formation of my own perspective (e.g., Borsboom, 2005, 2006; Sijtsma, 2009), even if my perspective may diverge from the authors' current perspective. In

Correspondence should be made to Daniel McNeish, Department of Psychology, Arizona State University, PO Box 871104, Tempe, AZ 85287, USA. Email: dmcneish@asu.edu

this way, my intention is to treat arguments in these previous sources at face value, independently of who made them.

In the remainder of this paper, I begin by summarizing some recent reviews of psychometric practices in empirical studies and the recent calls to improve these practices. Then, I consider how CTT-based sum scores align with these calls, with specific attention on (a) using sum scores to predict other variables, (b) using sum scores as an outcome in a subsequent analysis, and (c) how CTT-based sum scores may affect the ability to provide evidence that a particular construct is being captured with minimal assumptions intact. Lastly, I reflect on whether sum scoring—even if mathematically justified—is well-positioned to improve how empirical studies approach measurement and psychometrics and, ultimately, improve our understanding of behavioral phenomena.

Psychometric Practices in Empirical Studies

There have been growing concerns that methodological and statistical practices have adversely impacted the replicability and reproducibility of conclusions in empirical studies within the behavioral sciences (e.g., Nosek et al., 2022; Pashler & Wagenmakers, 2012; Tackett et al., 2019). Issues related to *p*-hacking, researcher degrees of freedom, or hypothesizing after results are known (HARKing) have received widespread attention and have triggered calls for reform of statistical practices (e.g., Rodgers & Shrout, 2018; Simmons et al., 2011; Wicherts et al., 2016). Among these examinations of empirical studies, researchers have recently noted that measurement and psychometrics may be an underappreciated source of replication issues (e.g., Flake & Fried, 2020; Schimmack, 2021; Soland et al., 2022a).

Specifically, scale scores are frequently used without accompanying evidence that scores are necessarily meaningful (e.g., that they capture an intended construct or predict a relevant outcome). For brevity, I refer to this as validity, though I acknowledge that there are different perspectives on the precise definition of validity. In the interest of maintaining a broad focus, I try to avoid those distinctions and proceed from the perspective of an empirical researcher trying to adhere to generally endorsed best practices rather than debating merits of different theoretical notions of what constitutes validity, especially because there appears to be general consensus across competing definitions that scores should mean something, irrespective of one's perspective on how that evidence should be acquired (e.g., Evers et al., 2012; 2015).

Among these review studies, Crutzen and Peters (2017) found that 2% of 288 reviewed health psychology scales provided validity evidence. Flake et al. (2017) reported that 21% of 177 social psychology studies reported validity evidence when using previously established scales and only 2% of 124 author-created scales were accompanied by validity evidence. Flake et al. (2017) also note that 19% of studies edited existing scales without validation of the revised scale. Weidman et al. (2017) reported that 69% of 356 scales in emotion research were newly created without validity evidence and Shaw et al. (2020) reported that 79% of 43 author-created scales provided no validity evidence. Higgins et al. (2023) reported that 14% of 925 studies using the popular Reading the Mind in the Eyes Test in clinical psychology cited previous validity evidence and 23% provided new validation evidence for their sample (37% total provide some validity evidence). Maassen et al. (2024) reviewed 918 scales in empirical psychology studies between 2018 and 2019 that had at least three items, sum scored, and compared scores across groups or time and found that only 4% considered measurement invariance (which is included as a component of validity in some definitions).

Potentially more problematic is that trends in these practices are largely unchanged from decades past. Qualls and Moss (1996) reviewed 2167 measures in APA journals in 1992 and found 32% reported validity evidence. Hogan and Agnello (2004) reviewed studies in APA journals between 1991 and 1995 and found 2% of the 696 studies reported validity evidence. Evers et

al. (2010) reviewed tests submitted to COTAN in the Netherlands across time and found that, between 1982 and 2009, the percentage of tests with good, sufficient, and insufficient validity evidence was mostly unchanged across time with only a modest improvement between 1992 and 2000 but no improvement between 2000 and 2009 (the COTAN validity benchmark changed in 1997, so these comparisons are approximate).

As summarized by Flake and Fried (2020), "The lack of information about measures is a critical problem that could stem from underreporting, ignorance, negligence, misrepresentation, or some combination of these factors. But regardless of why the information is missing, a lack of transparency undermines the validity of psychological science" (p. 457). As succinctly put by Brennan (2006), "validity theory is rich, but the practice of validation is often impoverished" (p. 8). Inattention to measurement and validity is especially pressing with the rise of new data structures like big or intensive longitudinal data and associated machine learning methods, which compound measurement deficiencies and magnify the 'garbage in, garbage out' principle of data analysis (e.g., Adjerid & Kelley, 2018; Bleidorn & Hopwood, 2019; Jacobucci & Grimm, 2020; Vogelsmeier et al., 2024).

Inadequate psychometric practices also impact replication efforts because it may be unclear if scores used as outcomes in empirical studies represent their intended construct or if they are merely capturing noise (Flake, 2021). Consequently, it can be unclear if failed replication efforts indicate that the theory being tested may not hold or if failed replication is attributable to measurement error (Loken & Gelman, 2017). Flake et al. (2022) note that fewer than 10% of replication efforts attempt to provide validity evidence for measures, which makes it difficult to differentiate between these two possibilities. To paraphrase Flake and Fried (2020), does *p*-hacking matter if scores are just noise to begin with? That is, statistical analysis is downstream of measurement, so if measurement is deficient, inferences may be inconclusive regardless of the quality and rigor of the ensuing statistical analysis.

Did Empirical Studies Ever Stop Using CTT?

A motivating premise of Sijtsma et al. (2024) is that sum scores have been banished to the psychometric mausoleum (p. 84) and that sum scores have lost ground quickly to IRT (p. 89). This is likely true in the context of operational psychometrics and in the methodological research literature, but CTT-based sum scores do not appear to be losing ground in empirical studies.

CTT has a lower barrier to entry and most empirical researchers have been using—and continue to use—CTT. For instance, Embretson (2004) wrote "the majority of psychological tests still were based on classical test theory" (p. 8). Wilson et al. (2006) note, "the CTT approach is by far the most widely known measurement approach, and, in many areas, is the most widely used for instrument development and quality control" (p. i20). When discussing psychometrics in medical research, Blanchin et al. (2011) noted, "To date, the choice of a statistical strategy for the analysis of such data is usually based on CTT rather than on IRT and seems to more likely rely on the researcher's practice and familiarity with CTT than on scientific grounds" (p. 826) and Gorter et al. (2016) write "despite the advantages of using IRT, in practice, sum scores are often used in the analysis" (p. 141). In a past Psychometric Society presidential address, Sijtsma (2012) wrote, "IRT is not the norm for test construction even though most psychometricians would prefer its use to the use of CTT" (p. 5).

In a review of industrial-organizational psychology, Foster et al. (2017) found that "in spite of the complementary nature of IRT and CTT, current research predominantly utilizes the latter" (p. 478). Foster et al. (2017) also found that—even in a psychometrically sophisticated empirical subfield like industrial-organizational psychology—only 31% of a survey of 343 industrial-organizational psychologists responded that they use IRT. Among those who do not, 45% said they believed classical methods worked fine and 21% said they never learned IRT. If empirical researchers are the intended recipient of the message that sum scores based on CTT represent the

field's greatest accomplishment, they may not need additional convincing because CTT already appears to be a popular and primary choice among this audience.

CTT, Sum Scoring, and Reporting Practices

Preferences for latent variable models in the psychometric literature have not necessarily translated to empirical studies and sum scores motivated by CTT do not appear to have gone—or be going—anywhere. However, because CTT delegates central tasks like dimensionality and invariance assessment to latent variable methods (e.g., Sijtsma et al., 2024, p. 100), empirical analyses equipped only with CTT tend to be incomplete by current standards, which possibly contributes to the state of psychometric reporting practices reported in the previous subsection where empirical studies simply do not engage with or report tasks outside the direct purview of CTT.

Sijtsma et al. (2024) note in their discussion that, "it is important to notice that researchers do their best to assemble item sets they believe to share the common core of the attribute of interest. They use psychometric methods such as corrected item-total correlations, principal component analysis, and FA [factor analysis] to assess the homogeneity of their experimental item sets before estimating the sum score's reliability and other psychometric properties." (p. 106). I do not dispute that researchers are doing their best and do not believe that poor psychometric practices reported in review studies are out of malice; my presumption would be lack of training (e.g., Aiken et al., 2008; Howard, 2024). However, reviews of psychometric practices in empirical studies do not support this statement, which seems to overestimate the psychometric sophistication underlying sum scoring in empirical studies.

A researcher who uses psychometric methods prior to summing responses and calculating reliability would seem to be in the minority. In fact, Flake et al. (2017) found that 18% of studies did not report reliability nor validity information, which mirrors the lack of reliability reporting in 19% of studies reported by Crutzen and Peters (2017). These values also track the 19% of North American psychology PhD programs in Aiken et al. (2008) who reported that their graduates were not equipped to perform a reliability assessment. So, a nontrivial proportion of studies do not consider validity or reliability information before or after summing item responses to create scores.

This prompts my reticence to broadly embracing sum scores for empirical studies—if sum scoring were synonymous with thoughtfully weighing the merits of different approaches and opting for CTT's mathematical elegance or simplified interpretation, I would be perfectly content and would have no objections (e.g., Stochl et al., 2022 provide an exemplary analysis supporting sum scoring). However, this is rarely the case and sum scoring more often is an ad hoc procedure—possibly without considering reliability and probably without considering validity—that is propped up by CTT because it is "a commonly accepted escape route to avoid notorious problems in psychological testing" (Borsboom 2005 p. 47).

Prescriptively, there are sound mathematical arguments to support sum scoring, but descriptively, most researchers are not appealing to any of them and chose to sum score because it is simple, intuitive, and widely accepted (possibly because reviewers and editors are doing the same thing themselves). Frequently, sum scoring is not a step supported by a broader psychometric plan so much as it is the entirety of the psychometric plan.

The situation feels analogous to the underpants gnomes in *South Park* (1998), who have a three-phase business plan where Phase 1 is to collect underpants and Phase 3 is profit. The joke is that they cannot figure out the second phase that connects Phase 1 and 3. Applied psychometrics seems to follow a similar plan where Phase 1 is collect item responses and Phase 3 is to compare people, but there is not always thought or planning dedicated to the second phase (perhaps giving new meaning to (g)nomothetic span; Embretson, 1983). Based on reviews of empirical studies, sum scores often serve as a means to an end and a path of least psychometric resistance to advance

from Phase 1 to Phase 3, which employs CTT not so much as a motive than as a convenient retroactive absolution.

As discussed toward the end of the current paper, elevating the status of the sum score may unintentionally preserve psychometric illiteracy among empirical researchers whose penchant for sum scores is not informed by potential merits of CTT (or by any psychometric theory) but rather by lack of methodological training, unawareness of alternative methods, or lack of motivation to engage with rigorous psychometric analysis. If researchers sum score based on the reasoning provided in Sijtsma et al. (2024), that would undoubtedly be a benefit for empirical studies. However, if empirical researchers interpret the superlative title of Sijtsma et al. (2024) as an endorsement of typical practice, psychometrics might have another obstacle in escaping from the periphery of empirical researchers' minds because principled approaches to sum scoring and how researchers currently approach sum scoring look rather different.

Of course, the information from these review studies could be interpreted in different ways. One reaction consistent with Sijtsma et al. (2024) is that psychometric practice in empirical studies is so poor that there is a benefit to simplification with sum scores and their milder assumptions because there are already enough problems without complex psychometric models. My reaction is that the poor state of psychometric practice in empirical studies is an opportunity for improvement and that endorsing sum scoring may not help disrupt the bleak state of psychometric practices. The sections that follow provide some rationale for my perspective.

Prediction with Sum Scores

Stochastic Ordering

Sijtsma et al. (2024) provide a simulation showing a set of conditions where the sum of binary items predicts a single true underlying construct as good or better than scores estimated from latent variable models, especially in the likely event where the item response function is not precisely known (also see Hopwood & Donnellen, 2010 for related arguments about benefits of prediction). The take-home message that followed is that sum scores derive their value from predicting external behavior and—because sum scores stochastically order an underlying attribute such that higher sum scores are associated with higher latent variable scores, on average—sum scores can have similar predictive ability as an estimated latent variable score (p. 106).

As one practical consideration, the stochastic ordering principle is upheld with unidimensional constructs informed by items with binary responses; however, reviews of empirical studies find that most researchers are not using binary response formats or unidimensional constructs. For instance, Flake et al. (2017) reported that 81% of empirical studies collected responses from ordinal Likert-type scales, whereas only 4% used binary response formats. Similarly, Maassen et al. (2024) reported that 5% of studies reported binary response scales and 95% used three or more response options. Jackson et al. (2009) reviewed 1409 factor analyses in psychology journals interested in scale development, construct validation, or measurement modeling and reported that only 9.5% were unidimensional, whereas 73.1% were multidimensional (the remaining percentage focused on models for invariance or multiple groups) while also finding that "the overwhelming majority" of studies used Likert-type items and treated them as continuous (p. 18).

Sums scores do not stochastically order a latent attribute with ordinal items (Borsboom, 2005, p. 124; Hemker et al., 1996, 1997), so this property will not necessarily hold in many empirical studies. As noted by Sijtsma et al. (2024), the impact tends to be mild such that the latent variable will still be approximately ordered by sum scores, on average, though the distortion does appear to increase with fewer items and more response options (van der Ark, 2005); the typical number of items per construct in empirical studies tends to be somewhat small (about 7

in empirical psychology studies; Flake et al., 2017; Jackson et al., 2009; Maassen et al., 2024). More importantly, stochastic ordering breaks down quickly in the presence of multidimensionality (Borsboom, 2005, pp. 123–124) because correlations between latent variables can rapidly erode correspondence between sum and latent variable scores (Pruzek & Frederick, 1978, p. 262) because sum scores tend to have difficulty incorporating correlations between constructs in the scoring process.

Stochastic ordering supports similar predictive ability of sum scores and latent variable scores in certain contexts (e.g., a bivariate correlation of summed binary responses and an underlying unidimensional construct), but these contexts do necessarily not align with what empirical researchers often possess (e.g., ordinal responses and multidimensionality). It might be hasty to extrapolate the predictive performance of sum scores from a simulation using binary unidimensional data as a general property of sum scores on account of stochastic ordering given that this property is known to hold in circumstances that may be uncommon in empirical data.

Additionally, it is relevant to note that stochastic ordering is about ranking the *expected value* of the latent variable score rather than ranking latent variable values of specific *individuals*, so prediction may be affected depending on the target of inference. This will be discussed next.

Moderated Associations, Individual Prediction, and Heterogeneity

Prediction often extends beyond bivariate correlations and can include nonlinear or moderated relations among multiple variables. Latent variable models can incorporate moderating characteristics into scores (a difficult task for CTT-based scores) to improve predictive ability for individuals when information about rank order of latent variable expected values is inadequate.

The moderated nonlinear factor analysis model (MNLFA, Bauer, 2017; Curran et al., 2014) is one example that allows all item parameters to be potentially moderated by discrete or continuous variables. This model has been shown to improve predictions over sum scores in both simulated data (Curran et al., 2016, 2018; Gottfredson et al., 2019) and empirical clinical diagnosis data (Coxe & Sibley, 2023; Hussong et al., 2019; Morgan-López et al., 2023; Soland et al., 2022b) .

Differences are not trivial—Morgan-López et al. (2023) meta-analyzed 25 post-traumatic stress disorder (PTSD) studies with item-level data and found that PTSD diagnostic concordance for individuals' PTSD was 73% with sum scores but 93% with MNLFA scores. If fixing diagnostic specificity to 80%, sensitivity was 48% with sum scores but 89% with MNLFA scores. This may occur because, even in conditions when stochastic ordering holds, it only speaks to expectations (i.e., people with higher sum scores—on average—have a larger ability than people with a lower sum score), so stochastic ordering is not necessarily satisfactory for predicting or classifying individuals (Zwitser & Maris, 2016) . Individuals are more likely to be the target of inference in empirical subfields that tend to emphasize psychometrics and scale scores like education and clinical psychology (Speelman et al., 2024), as is the case of predicting an individual PTSD diagnosis in Morgan-López et al. (2023).

Perhaps this demonstrates the point in Sijtsma et al. (2024) that something as simple as a sum score retains remarkably high predictive ability relative to complex models that demand far more computational resources. I also do not want to discount the nontrivial advantages of CTT-based sum scores when sample size is small or possibly when sample size is very large (e.g., where computational demand becomes excessive) because complex models encounter far more problems as a function of sample size at either extreme, whereas CTT-based sum scores scale easily across the entire sample size distribution. Points about potential model misspecification are also insightful as latent variable models are not always specified carefully and there are discernible considerations for matching scoring procedures to the appropriate level of rigor and the audience interpreting scores (e.g., classroom tests are fairly low stakes and less sophisticated scoring methods suffice and are easier to interpret even if they may not be ideal).

Nonetheless, if the target benchmark is prediction in a scientific study, there are several contexts where augmented latent variable models or machine learning methods can have greater predictive accuracy than a sum score (Gonzalez, 2021; also see Tay et al., 2020 for discussion of validity of scores derived from machine learning), especially if emphasizing prediction and removing the requirements that a score needs to capture a specific construct or be easily interpretable by a lay audience because criticisms of model complexity are less pertinent. For instance, we have no idea if the scoring model in Morgan-López et al. (2023) is correctly specified, but if we only care about prediction, it does not matter because the scores it produces considerably outperform an unweighted sum of PTSD symptoms.

Directly Using Items as Predictors

Outside of latent variable modeling, there is an emerging literature on the predictive benefit of using individual items as predictors over a sum of items (e.g., Donnellan et al., 2023; Fried & Nesse, 2014; McClure et al., 2024; Müller et al., 2023; Revelle, 2024). For instance, McClure et al. (2024) found that when using Beck's Depression Inventory II to predict suicidal ideation, using the sum score as a predictor yielded an R^2 of 0.20 but using individual items as predictors had an R^2 of 0.38. And when predicting suicidal ideation from the Patient Health Questionnaire-9 (PHQ-9), the R^2 using the sum score as a predictor was 0.39 versus an R^2 of 0.58 when using individual items as predictors.

Müller et al. (2023) found comparable results when using individual or summed personality disorder criteria for predicting several different outcome variables. In Müller et al. (2023), predictive performance was especially different when comparing prediction using individual criteria to prediction using a sum of criteria across all syndromes (versus sums of criteria intended to represent a single syndrome). Rather than creating a composite predictor by summing item responses according to a predefined weighting scheme (often equal weighting), using the item responses as predictors directly can permit heterogeneous predictive contributions of different items (Fried, 2015; Fried & Nesse, 2015) without requiring assumptions about dimensionality or assuming that a single construct underlies item responses. If the objective is predictive accuracy with minimal assumptions about scores, using individual items as predictors may be a more attractive option than a sum.

Sampling Variability

To reorient to a topic that is related to the previous subsection but that is not directly related arguments in Sijtsma et al. (2024), preference for sum scores is sometimes based on arguments that they are consistent across studies, whereas estimated scores from latent variable models are built upon parameter estimates that have sampling variability (e.g., Russell, 2002). Wainer (1976) generally made this argument by showing that loss of predictive accuracy in regression is minimal if regression coefficients associated with standardized predictors are replaced with $+1$, 0, or -1 (see also, Cohen, 1990, p. 1306). This argument is often extended to measurement models to support replacing estimated weights from factor analysis with equal weights as in a sum score.

However, Pruzek and Frederick (1978) showed that some assumptions made by Wainer (1976) in the regression context (e.g., predictors are uncorrelated; standardized weights are uniformly distributed over [.25, .75]) may not readily extend to measurement models. Whereas predictors in linear regression explain a fixed amount of variance in a single outcome, measurement models are multivariate such that each item has a separate amount of variance that can be explained by a latent variable. Pruzek and Frederick (1978) note that this affects the tenability of assumptions upon which arguments for equivalent predictive accuracy with equal weights are based. They show examples where there can be meaningful loss in predictive accuracy when estimated weights are replaced with equal weights. Loss of accuracy will not necessarily occur (e.g., if the range of standardized coefficients is limited), but conditions encountered in factor analysis are more susceptible to loss than linear regression.

Somewhat ironically, replacing estimated coefficients with equal weights is not commonly practiced in regression where it has stronger support for retaining equivalent predictive accuracy. However, replacing estimated coefficients with equal weights is more common in measurement contexts despite somewhat less support that resulting scores will be comparable (e.g., that the calculated scores will relate comparably to the true construct).

Concern about sampling variability is legitimate, but these concerns can be selectively applied such that sampling variability is sometimes used to justify equally weighted sum scores, only for researchers to proceed to a prediction stage where these sum scores are used in a regression model with uniquely estimated coefficients such that sampling variability of regression coefficient estimates is suddenly no longer a concern. If sampling variability is concerning, why estimate regression weights in the prediction model instead of constraining them to be equal or to predefined values to avoid sampling variability as in the scoring model?

Concerns about sampling variability and out-of-sample performance in measurement models may also be mitigated by more recently developed methods like as regularization (e.g., Huang, 2022; Jacobucci et al., 2016; Li & Jacobucci, 2022; Liang & Jacobucci, 2020), incorporating sampling variability into scoring (Tsutakawa & Johnson, 1990) , and fixing weights to values from a previous validation (Kim, 2006; König et al., 2021) .

Prediction is meaningful for contexts where scores are used as independent variables, but it may not always be as useful in situations where the scores are intended to be an outcome in a subsequent analysis. The next section reviews the literature on using sum scores as an outcome in empirical analyses.

Sum Scores as Dependent Variables

The stochastic ordering property of sum scores is remarkable for its simplicity, but the fact that it yields scores on an ordinal scale potentially limits performance if scores are subsequently used in analyses where the interest is quantifying group differences or change over time rather than bivariate correlations (Reise & Henson, 2003) . It is important to note that high correlations between sum scores and latent variable scores do not imply equivalence of subsequent performance (McNeish, 2023a) because Pearson correlations are largely insensitive to monotonic transformations (Reise & Waller, 2009) . Altman and Bland (1983) emphasize that high correlations between two methods do not imply agreement between two methods, which was demonstrated by Gonzalez et al. (2021) who showed that two scores correlating .998 could still have meaningfully different correlations with a third variable. So, the common finding that sum scores and latent variable scores are highly correlated does not guarantee that they will have interchangeable performance or conclusions when using different types of scores as an outcome in subsequent analyses.

Previous studies have reported poorer performance of sum scores to detect underlying effects, trends, or group differences in different modeling contexts like regression discontinuity (Soland et al., 2023) , growth modeling (Edwards & Wirth, 2009; Edwards & Soland, 2024; Fraley et al., 2000; Gorter et al., 2020; Kuhfeld & Soland, 2022; Luningham et al., 2017; Proust-Lima et al., 2019; Tang et al., 2023) , randomized or clinical trials (Gorter et al., 2016; Kuhfeld & Soland, 2023; Kessels et al., 2021; Soland, 2022) , machine learning (Gonzalez, 2021; Jacobucci & Grimm, 2020) , time-series and intensive longitudinal analysis (Vogelsmeier et al. 2019, 2021, 2022), growth mixture modeling (Soland et al., 2024) , and partial least squares for formative latent variables (Hair et al., 2024) .

Ramsay and Wiberg (2017) note that sum scores in some application areas can congregate at extreme scale points and form floor or ceiling effects (e.g., Pelt et al., 2023; Schwabe & Van den Berg, 2014; Van den Oord et al., 2003) and alternative scoring methods—even under stochastic

ordering—can improve prediction through better scaling (also see, Proust-Lima et al., 2011; Van den Oord & Van der Ark, 1997). Relatedly, Liu and Wang (2021) found that a majority of studies in flagship education and psychology journals using *t*-tests (57%) and ANOVA (70%) have unaccounted floor or ceiling effects in their outcome variable, which can emerge from reliance on sum scores as dependent variables and can result in distorted inferences.

Maxwell and Delaney (1985) found that monotonic transformations of underlying latent variable scores (i.e., how sum scores relate to a latent variable scores) were insufficient for *t*-tests to be accurate, and Wilcoxon rank-sum tests were needed (i.e., sum scores need to be treated as ordinal in subsequent analyses, a rare occurrence in practice). Maassen et al. (2024) mention that group comparisons with sum scores may also be complicated by possible invariance (see also, Slof-Op't Landt et al., 2009), especially because researchers infrequently evaluate invariance, possibly because CTT does not have a strong framework for invariance assessment (e.g., Wilson et al., 2006).

Several studies have shown that using sum scores as an outcome in the common context of models with interaction terms such as factorial ANOVA or moderated multiple regression worsens performance (Embretson, 1996; Kang & Waller, 2005; Morse et al., 2012; Murray et al., 2016). This also applies when a model features an interaction of two sum scores as predictors and their measurement error is not accounted for (Hsiao et al., 2018). Note that Sum scores being compared in these studies are often raw sums of responses, but performance can be improved with transformation of sum scores (e.g., Murray et al., 2016) and arguments in Sijtsma et al. (2024) are compatible with transformed sum scores.

The main point is that the extending the stochastic ordering property of sum scores beyond bivariate correlations does not necessarily translate into accurate conclusions about underlying associations if the intent is for sum scores to be included as an outcome in a subsequent statistical model. When sum scores are used as outcomes, traditional models expect interval data where spacing and variability matter for proper inference, aspects which may not be preserved by sum scores that maintain average rank ordering (ordinal responses and multidimensionality for which stochastic ordering does not exactly hold can amplify differences in performance).

Sijtsma et al. (2024) present an interesting case of network models. Though I am admittedly not well-versed in this area, my basic understanding is that the network itself is typically the main interest as opposed to latent variable models where the interest is often to understand some latent structure so that scores can be calculated and passed onto the next stage of analysis or use. Stochastic ordering may not be sufficient when scores are passed on to models where spacing and variability are important. However, when the network itself is the focal interest, sum scoring may be more attractive because the stochastic ordering is much more appealing than it may be in contexts where a measurement model serves essentially as a preprocessing step to a subsequent focal analysis.

Unlike when scores are used as a predictor, sum scores used as outcomes (outside of network models) more often implicitly convey that the score represents a specific construct. It is often prudent to provide evidence that a score is accurately capturing the intended construct prior to using the score in a subsequent statistical model. Many of the studies cited in this section concerned sums of items that were known to come have a single underlying construct (e.g., because the data were simulated). However, empirical studies must gather evidence to establish a link between scores and a specific construct prior to using scores as an outcome. Considerations during this process are discussed in the next section.

Scores Intended to Capture a Specific Construct

Sijtsma et al. (2024) emphasize that CTT does not necessarily intended to capture a specific construct by saying: "the mathematics of [CTT] are independent of how one wishes to interpret the model …CTT is a truly minimal model in terms of assumptions…the crucial insight is that CTT can operate in the absence of assumptions regarding the test's dimensionality or factorial composition" (p. 100). Essentially, CTT guarantees the right to create composite scores by summing but does not guarantee that the resulting score will necessarily correspond to a specific desired construct or any construct at all.

The focus on correlation and external prediction with sum scores therefore makes inherent sense—CTT does not necessarily specify a specific underlying construct, so the utility of CTT-based scores (like sum scores) can be derived from the extent to which they relate to or predict a relevant external target (e.g., Kane, 2006). Nonetheless, a common intention is for scores to represent a particular construct (e.g., Sijtsma et al., 2024, p. 106), which requires some evidence to establish a connection between item responses and the intended construct.

Sijtsma et al. (2024) emphasize that CTT does not exclude dimensionality restrictions whereby a certain construct may underlie responses (p. 87) but that "CTT does not allow the assessment of dimensionality simply because this is not part of CTT and that, therefore, researchers need to use dimensionality assessment methods from outside of CTT" (p. 100). This is helpful, but potentially introduces a conflict between maintaining CTT's minimal assumptions and employing methods outside CTT to demonstrate whether scores plausibly represent a particular construct.

For instance, if factor analysis were applied to extract evidence that a particular construct was underlying the item responses, several additional assumptions would be needed—from the list of assumptions not made by CTT (Sijtsma et al., 2024, p. 99), this would include Item 4 (scores do not need to satisfy a dimensional model), Item 5 (scores do not need to reflect the same attribute), possibly Item 6 (errors do not need to be independent), Item 7 (there are no distributional assumptions), and possibly Item 8 (error variances are the same for every person).

These assumptions are indeed unnecessary to create the scores or to assess reliability of scores, but they are needed to provide crucial support from a method like factor analysis that scores are a plausible representation of a particular construct. Perhaps it is not entirely accurate to attribute these additional assumptions to CTT directly since they are technically made by a supporting method like factor analysis, but it also does not seem entirely accurate to assert that CTT necessarily makes minimal assumptions if its use in some contexts depends on an accompanying assumption-laden method. In other words, does CTT embody the assumptions of the methods used to justify it? How does dependence between CTT and the method used to justify it impact CTT assumptions?

More broadly, there appears to be a distinction between the purest version of CTT and the dimensionality-restricted version of CTT that many researchers seek when they want to interpret scores as reflecting a particular construct. Defense of CTT and CTT-based sum scores is appropriately quick to highlight the minimal assumptions under which scores can be created and reliability can be defined. However, a qualification is that the purest version of CTT is agnostic to what the score actually captures and is impervious to quantitatively evaluating aspects like dimensionality because the purest form of CTT "has no room for the important and challenging psychometric question of how theoretical attributes are related to observations" (Borsboom, 2006, p. 430). CTT does not exclude such examinations, but it does not necessarily require or encourage them either. Omission of validity aspects from defenses of CTT-based sum scoring is therefore not missing at random—CTT was not built to accommodate validation efforts because CTT does not concede that a theoretical attribute or construct exists.

Bringing dimensionality restrictions into CTT comes at the expense of additional assumptions—there is a trade-off between minimal assumptions and scores representing something specific. As

noted by Borsboom and Mellenbergh (2004), "the classical test theory model is largely untestable unless auxiliary assumptions, such as equal error variances across subjects, are imposed, and it is certainly never tested in actual research." (p. 108). Anecdotally, the motivation of McNeish and Wolf (2020) was to outline a method to facilitate testing whether dimensionality-restricted sum scores justifiably capture a single construct in empirical research, which requires additional assumptions beyond those required for the purest form of CTT. I later came across Beauducel and Leue (2013), which follows the same theme but suggests a different set of constraints that correspond to justifying a dimensionality-restricted sum score (also see Rose et al., 2019 for a third alternative specification of a similar idea).

There are arguments that a latent variable model is a type of Wittgenstein's ladder such that its purpose is to justify a sum score, but—after having done that—the model is no longer useful and need not inform each person's value or position on the construct.[1] There is merit to this idea and precedence for viewing a sum score as a coarse version of predicted construct score (e.g., Grice & Harris, 1998; Grice, 2001). That is, once evidence for dimensionality is obtained, parameter estimates from a latent variable model used in dimensionality assessment can be discarded for scoring.

A possible counterargument may be that some prevailing definitions of validity consider it as a property of scores (e.g., AERA, APA, & NCME, 2014). There are several methods to predict scores for unobservable constructs (e.g., DiStefano et al., 2019), but not all methods imply the same reproduced interitem covariance matrix (Beauducel, 2007; Beauducel & Hilger, 2020), whose comparison to the observed interitem covariance matrix serves as the basis of factor analytic fit. Using the labels from Grice (2001), "refined" methods that incorporate estimated factor loadings for weighting scores tend to have equivalent reproduced matrices (Beauducel, 2007) , but "coarse" scores that use a simplified weighting scheme like an unweighted sum do not (Beauducel & Hilger, 2020; Beauducel & Leue, 2013).

From this validity perspective, the argument is that the scoring model and the factor model are not necessarily independent and changing the scoring method can change validity evidence gleaned from factor analytic fit (e.g., Embretson, 2007, p. 453; Thissen, 1983, p. 215). If validity is considered a property of scores, scores based on different weighting schemes—even from the same factor structure—may not necessarily have interchangeable validity evidence. As put by Edwards and Wirth (2009): "there is indeed something odd about the common practice of using factor analysis to establish the dimensionality of a scale but then ignoring the parameter estimates themselves when creating scale scores. Statements about the adequacy of a model from a factor analytic standpoint may not apply when the parameters from that model are ignored." (pp. 84–85; also see Schreiber, 2021, p. 1009).

As convincingly argued in Sijtsma et al. (2024), deferring to latent variable models to justify sum scores provides sufficient—but not necessary—conditions for sum scoring because applying a latent variable model applies a more restrictive set of assumptions than required by the purest form of CTT. This argument makes complete sense if there is no intent for scores to represent a specific construct. However, the pure form of CTT has so few assumptions *because* it is indifferent to what the scores represent and because constructs are not represented in the model. Validity and justification of scores is therefore not an essential feature.

If moving to a dimensionality-restricted version of CTT where constructs are presumed to be present and validity is more consequential, it becomes difficult to maintain minimal assumptions and avoid moving toward latent variable conceptualizations. As put by Borsboom (2005), "classical test theory does not formulate a serious account of measurement, and therefore is inadequate to deal with the question of validity. In fact, if it begins to formulate such an account, it invokes a kind

[1] I am indebted to Denny Borsboom for providing this information during review. I was not familiar with this idea in the initial version of the manuscript and, even if I were, I would not have been able to articulate it as clearly.

of embryonic latent variable model" (p. 144). Of course, when this was written, network models and component models were much less developed and much less visible in behavioral sciences, so alternatives to CTT have expanded in the intervening 20 years. Though these methods reduce reliance on the conventional reflective latent variable model like factor analysis, the general point that some other method must accompany dimensionality-restricted CTT remains relevant.

In short, minimal assumptions are great for assessing reliability but can be a detriment for assessing aspects of validity because the purest form of CTT is insulated from such evaluation. There seems to be a disconnect between the abstract notion of what mathematics permits and the practical context of what researchers are doing. The schism between mathematics and empirical applications is highlighted by Borsboom (2005), who wrote that CTT is "so enormously detached from common interpretations of psychological constructs, that the statistics based on it appear to have very little relevance for psychological measurement" (p. 47).

It is unclear how sum scores motivated by the purest form of CTT fit into recent calls for greater attention to validation and greater transparency in psychometric reporting because its definitions do not seem interested in or capable of addressing such questions without deferring to latent variable methods and the additional assumptions they impose.

Assumptions and Intent

This situation seems analogous to the role of assumptions in ordinary least squares to estimate parameters in linear regression. Say we have a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ where $\mathbf{y}$ is the outcome vector, $\mathbf{X}$ is a matrix of predictors, $\boldsymbol{\beta}$ is a regression coefficient vector, and $\mathbf{e}$ is a vector of errors. The solution for $\boldsymbol{\beta}$ that minimizes the squared differences between the observed data and predicted values is $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$, which only assumes no perfect collinearity (i.e., $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$ exists) and, for asymptotic consistency, exogeneity (i.e., $E\left(\mathbf{e}|\mathbf{X}\right) = 0$).

Regression lines can be fit through data without any of the typical assumptions taught in introductory statistics like independence, normality, or homoskedasticity. However, as soon as inferential evaluations of the model are of interest (e.g., whether a regression coefficient is 0 in the population), the trio of assumptions encompassed by $\mathbf{e} \sim N\left(0, \sigma^2\mathbf{I}\right)$ is needed. Regression lines can be *created* while maintaining few assumptions, but they cannot always be *evaluated* while maintaining few assumptions.

Though the motivation and context are different, the general idea seems related to CTT-based sum scores. According to CTT, sum scores can be *created* with few assumptions. However, construction of relevant statistical tests or procedures to *evaluate* aspects of scores other than reliability is difficult while maintaining few assumptions (e.g., dimensionality or invariance). Promoting CTT-based sum scores on the basis that the underlying mathematics do not require many assumptions is defensible, just like using linear regression via ordinary least squares for prediction is defensible without independence, normality, or homoskedasticity.

But these arguments are contextual. Just as minimal assumptions to create a regression line with ordinary least squares are unhelpful to researchers interested in inference (which is presumably why students are usually taught that ordinary least squares assumes $\mathbf{e} \sim N\left(0, \sigma^2\mathbf{I}\right)$), the minimal assumptions of CTT-based sum scores may not be helpful to empirical researchers who want to use scores to represent a particular construct because assumptions are the cost of interpreting scores in a specific way.

Validation Without Latent Variables

Sijtsma et al. (2024) make a strong case for the latent variable model as the justification for sum scores, so it is unsurprising that the minority of empirical studies that do report validity evidence tend to emphasize methods like factor analysis. For instance, recent reviews of validity reporting in empirical studies find that the percentage of studies presenting evidence based on factor analysis or internal structure is 90% (Shear and Zumbo 2014), 89% (Collie & Zumbo, 2014) , 92% (Gunnell et al., 2014), 85% (Chinni & Hubley, 2014), and 77% (Hubley et al., 2014) .

Nonetheless, researchers could work around reliance on latent variable models and their additional assumptions by employing methods from the content (e.g., Aiken, 1980; Mislevy et al., 2003; Sireci, 1998a, 1998b; Sireci & Faulkner-Bond, 2014) or response process families of validation methods (e.g., Embretson, 1983; Mislevy et al., 2002; Padilla & Benítez, 2014). These methods provide qualitative evidence that the items cover relevant aspects of the intended construct or that respondents are understanding the items consistent with the way an attribute is defined, respectively, and could conceivably be done without need to impose additional assumptions as in the case of quantitative approaches (though, ideally, quantitative and qualitative sources would be provided together for a more holistic validation).

Based on reviews of reporting practices, this type of validation evidence is exceedingly rare in empirical studies. For instance, a review of papers in the *Journal of Educational Psychology* between 2000 and 2010 by Collie and Zumbo (2014) found that 16% reported evidence of test content and 0% reported response process evidence. Hubley et al. (2014) found that of articles published in *Psychological Assessment* and the *European Journal of Psychological Assessment* between 2010 and 2012, 2% reported response process evidence and 0% reported evidence of test content. If minimal assumptions of CTT-based sum scores are a key advantage to retain even when the goal is to capture a specific construct, the importance of content and response process validation and how to collect such evidence may need to be better communicated to empirical researchers.

Will Sum Scoring Help Empirical Studies?

The properties and underlying mathematics of CTT and sum scores are interesting to contemplate, but the mathematical machinery underlying CTT and sum scores ultimately may be less relevant than what CTT and sum scoring can offer to empirical researchers. For instance, if IRT were developed first and CTT came afterward, would CTT be defended as vigorously based on the merits of what it offers or does the affinity for CTT in some part come from its historical context, the elevated status of those who initially conceived the idea, or that it is simply easier to apply? Borsboom (2006) presents this point articulately, saying:

> *In an alternative world, where classical test theory never was invented, the first thing a psychologist, who has proposed a measure for a theoretical attribute, would do is to spell out the nature and form of the relationship between the attribute and its putative measures. ... This would lead the researcher to start the whole process of research by constructing a psychometric model. After this, the question would arise which parts of the model structure can be tested empirically, and how this can best be done. Currently, however, this rarely happens. In fact, the procedure often runs in reverse.* (p. 429).

If (a) latent variable models provide the basis for sum scores that CTT itself does not provide (Sijtsma et al., 2024, p. 97), (b) latent variable models provide or imply validity evidence for sum scores that CTT does not provide (Sijtsma et al., 2024, p. 100), and (c) latent variable models allow additional assessments like differential item functioning and invariance in ways that CTT could not realize (Sijtsma et al., 2024, pp. 106–107); at what point do we consider the latent variable approach as a more complete framework for typical empirical settings, especially in light of recent research showing that using individual items often has greater predictive validity than a sum of item responses?

This dissonance emerges in one possible CTT-based workflow that satisfies modern psychometric standards like those from AERA, APA, and NCME (2014) where (a) scores are motivated by minimal assumptions of CTT, (b) scores are validated by conducting a factor analysis that

introduces several assumptions to evaluate dimensionality (and, possibly, invariance), and (c) the assumptions and parameter estimates from the factor analysis are disregarded and the sum score is used as an estimate of the underlying construct and its reliability is reported with coefficient alpha.

In such a case, is it still accurate to maintain that the scores are based on minimal assumptions if CTT leans on factor analysis or auxiliary assumptions to provide evidence that scores have the intended dimensionality? Why discard the information provided by the factor analysis about relative weights of the items (and upon which factor analytic fit may rely) in favor of a predefined weight scheme? If factor analysis and its assumptions are needed for the validation portion of the analysis, why not make these assumptions from the onset and operate entirely within a factor analysis framework given that it similarly has mechanisms for reliability estimation and scoring (and that scores created from factor analysis weights are equally or more reliable than scores created from unit weights; Hancock & Mueller, 2001; Li et al., 1996)?

CTT can only take the analysis so far before it must outsource remaining steps to another method. Dimensionality-restricted versions of CTT that are of interest in many empirical contexts where scores are intended to capture a specific construct seem to offer limited benefit over starting with factor analysis or IRT and building evidence for or against sum scores entirely within one of those frameworks. It seems equally fitting to describe this process as factor analysis with coarse factor scores as it does to describe it as CTT.

It would be odd to fit a regression model, use assumptions encompassed by $\mathbf{e} \sim N\left(0, \sigma^2\mathbf{I}\right)$ to compute standard errors for inferential testing, find that all predictors are plausibly non-null in the population, and then declare that the regression model did not assume $\mathbf{e} \sim N\left(0, \sigma^2\mathbf{I}\right)$ because the initial regression line created without assumptions was upheld. Similarly—at least to me—it seems odd to create sum scores based on CTT, use assumptions of factor analysis to assess dimensionality, find that a unidimensional model is plausible based on some criterion like RMSEA $< .06$, and then declare that the scores rely on minimal assumptions.

Relatedly, in a regression context, it would feel unconventional to fit a model with uniquely estimated coefficients and use its R^2 to describe the predictive ability or fit of a different model whose coefficients are constrained to be equal. However, in psychometrics, it is routine to use factor analytic fit from model with unequal weights as evidence for dimensionality of equally weighted sum scores even, though their model-implied covariance matrices may not be identical.

Again, there are no issues with applying constraints to avoid sampling variability or to prioritize consistency between studies, but aspects of the bias-variance trade-off are at play. In regression, using a penalized method like lasso reduces the sampling variability by reigning the coefficients toward zero, but a side effect is that R^2 typically decreases because the price for reducing sampling variance is an increase in bias (i.e., to obtain estimates with smaller between-sample variability, predicted values are slightly further for observed values in the data). In a measurement context, many are happy to accept the lower between-sample variance associated with equal weights, but they are not as keen to embrace the associated price that model fit is slightly worse because scores are a little less closely related to the construct (higher bias).

Of course, latent variable models have their own weaknesses that do not make them universally appropriate. I would be one of the first people in line to criticize how factor analytic fit is evaluated (e.g., McNeish, 2023b, 2023c), latent variable models tend to be accompanied by little substantive theory that can hamper their utility (e.g., Fried, 2020; Eronen & Bringmann, 2021), and latent variable models encourage overemphasizing quantitative components of validity (e.g., Alexandrova & Haybron, 2016; Peters & Crutzen, 2024; Wolf, 2023). Any method applied without purpose and thought will have deficiencies and replacing uncritical sum scoring with uncritical use of factor analysis or IRT will do little to remedy current psychometric issues in empirical studies. I realize that much of this text defends factor analysis, but I am in no way convinced that factor analysis should serve as the go-to method, and it has many problems that other methods

can circumvent. Whereas the psychometric literature has historically pitted CTT against reflective latent variable models like factor analysis, the breadth of options is expanding with recent work on the often-overlooked area of formative constructs and component modeling (Hwang et al., 2021; Rhemtulla et al., 2020; also see Hair et al., 2024 for a discussion of sum scoring formative constructs) and network models offer a complementary way to assess interitem covariances (e.g., Christensen et al., 2020; Epskamp et al., 2017). The broader point is that entering an analysis having decided to sum score and using other methods to work backward and justify summing seems to negate possible insight and contributions that could be garnered by starting with these methods from square one and building evidence for a particular scoring method or underlying structure.

If focusing on the practical issue of improving psychometric practices in empirical studies to improve our collective knowledge about behavioral phenomena, defaulting to CTT-based sum scores does not seem like an entirely effective strategy because CTT is not self-contained and its precise niche in modern psychometric analysis is somewhat unclear. There is not a complete framework for validity without exporting steps of the analysis. Sum score prediction beyond bivariate correlations can often be worse than simply using the individual items as predictors (which similarly requires few assumptions). CTT excels at reliability estimation if the score does not necessarily need to capture a specific construct, but if a specific construct is desired and its relation to item response is thought to be reflective, factor analysis and its assumptions that are already summoned for validation can also evaluate reliability.

CTT-based sum scores are not worthless, but the scope of where they excel or situations in which they are an optimal approach seems rather narrow relative to the interests of many empirical studies. With continuing expansion and refinement of alternative methods, there is more opportunity than ever to strive toward deeper understanding of psychological processes behind item responses or patterns of multiple item responses. Embracing sum scores and potential agnosticism about what a score captures concomitant with CTT might limit engagement with new approaches that can provide original conceptualizations of behavioral phenomena. Overreliance on the appeal and simplicity of sum scoring is partially responsible for some of the deficient measurement practices common in empirical studies given that summing responses has been the dominant approach in empirical studies for quite some time. Continuing down the same path—even with renewed mathematical justification—seems like it will maintain the status quo.

Final Remarks

Borsboom (2006) has a great closing line: “The current practice of psychological measurement is largely based on outdated psychometric techniques... I suggest we work as hard as possible to facilitate the emergence of a new generation of researchers who are not afraid to confront the measurement problem in psychology.” (p.438). I understand the intent of Sijtsma et al. (2024) was to vouch for stronger balance in psychometric approaches and appreciation for classical methods because the psychometric literature is dominated by latent variables and often beats up on CTT (perhaps unfairly) to motivate novelty of methods. For readers whose focus is psychometrics, Sijtsma et al. (2024) accomplish their task with exceptional clarity.

Conversely, for empirical researchers who are content to stay within the confines of CTT, influential references from luminaries in the field declaring sum scoring as the greatest accomplishment in psychometrics may inadvertently foster—rather than confront—continued dominance of classical approaches in scenarios where more modern approaches are advantageous or even necessary to complement CTT and justify sum scoring. The net effect may be further lowered motivation to learn or consider modern methods if the perception is that existing psychometric practice—which is often based on uncritical use of sum scores—is suitable.

Whenever there is momentum to move empirical researchers away from classic methods that dominate in empirical studies, the pendulum seems to forcefully swing back to defend CTT and CTT-based scores, leaving empirical researchers incapacitated and unsure how to proceed. Meanwhile, psychometricians and quantitative psychologists act surprised every time a new review paper shows that (a) empirical researchers do not seriously engage with psychometrics, (b) empirical researchers do not think learning modern psychometric methods is worthwhile, and (c) trends in desirable psychometric practices in empirical studies have been flat for decades.

Despite the latent-variable-heavy state of the psychometric literature, most empirical researchers never stopped using sum scores and they often do not accompany their sum scores with relevant validity evidence when scores are intended to capture a specific construct. There are interesting and meritorious mathematical arguments for sum scoring under minimal assumptions, but the contexts in which many empirical researchers are working (e.g., Likert responses, multiple constructs, intent for scores to represent specific constructs) can be orthogonal to the contexts under which these mathematical arguments optimally apply and mathematical arguments supporting CTT-based sum scoring do little to promote the importance of lacking practices like validity assessment that affect dependability of empirical studies. To be clear, I do not think sum scores are universally bad and it is entirely possible to build strong psychometric cases for sum scoring. However, I also think that uncritical use of sum scores by uninitiated empirical researchers undeservedly receives a pass partly on the basis of CTT.

My apprehension is that well-intentioned defenses of sum scoring will be interpreted by empirical researchers as reassurance to continue to avoid engaging with serious psychometric endeavors because the perceived message may be that the status quo is easy and sufficient. Sum scores can certainly be defended, but many instances of sum scoring in empirical studies are motivated by the simplicity of sum scores rather than any psychometric theory, evidence, or arguments. Of course, it is plausible that psychometrics really is as simple as summing responses, and I am the naïvely optimistic one who merely wants modern methods to be better to give our field credence and to show the empirical researchers, biostatisticians, and econometricians that psychometrics did not peak in the 1960s and that we have something meaningful to contribute to scientific discourse.

Nonetheless, psychometricians and quantitative psychologists could benefit from changing the objective function that we seek to maximize with our work. Rather emphasizing what mathematics might allow, we can better frame our arguments to (a) help empirical researchers understand how psychometrics can improve understanding of behavioral phenomena and (b) be more cognizant of challenges facing empirical researchers by meeting them where they are.

To adapt a line from Angrist (2004), psychometrics is too important to be left entirely to psychometricians (p. 201). At its core, psychometrics is an inherently applied discipline, and scores are the foundational unit of analysis in many subfields of behavioral science. Reviews of empirical studies find that (a) sum scoring still dominates, (b) the importance of validity is rarely embraced, and (c) little thought is generally put into creating scores despite their central role in subsequent analyses. Reinforcing a commonly applied approach will likely result in more of the same and seems unlikely to curb deficient psychometric practices in empirical studies.

Defense and support of CTT and CTT-based scores is a legitimate mathematically justified position for psychometricians who can appreciate nuances and who are comfortable working at a certain level of abstraction. However, CTT and CTT-based sum scoring was not designed with validity in mind, and potentially make validation less approachable to empirical researchers who are already struggling to provide validity evidence for their scores. Ultimately, approval for CTT-based sum scoring by psychometricians may be misconstrued by—and unhelpful for—empirical researchers who are on the front lines of behavioral research because "few, if any, researchers in

psychology conceive of psychological constructs in a way that would justify the use of classical test theory as an appropriate measurement model" (Borsboom, 2005, p. 47).

Declarations

Conflict of interest The author did not receive support from any organization for the submitted work, and the author has no financial or nonfinancial conflict of interest to disclose.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist, 73*(7), 899–917.

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63*(1), 32–50.

Alexandrova, A., & Haybron, D. M. (2016). Is construct validation valid? *Philosophy of Science, 83*(5), 1098–1109.

Altman, D. G., & Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society, Series D: The Statistician, 32*(3), 307–317.

Angrist, J. D. (2004). American education research changes tack. *Oxford Review of Economic Policy, 20*(2), 198–212.

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

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement, 40*(4), 955–959.

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods, 22*(3), 507–526.

Beauducel, A., & Hilger, N. (2020). On the fit of models implied by unit-weighted scales. *Communications in Statistics-Simulation and Computation, 49*(11), 3054–3064.

Beauducel, A., & Leue, A. (2013). Unit-weighted scales imply models that should be tested! *Practical Assessment, Research & Evaluation, 18*(1), 1–7.

Beauducel, A. (2007). In spite of indeterminacy many common factor score estimates yield an identical reproduced covariance matrix. *Psychometrika, 72*(3), 437–441.

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review, 23*(2), 190–203.

Borsboom, D., & Mellenbergh, G. J. (2004). Why psychometrics is not pathological: A comment on Michell. *Theory & Psychology, 14*(1), 105–120.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425–440.

Blanchin, M., Hardouin, J. B., Neel, T. L., Kubis, G., Blanchard, C., Mirallié, E., & Sébille, V. (2011). Comparison of CTT and Rasch-based approaches for the analysis of longitudinal patient reported outcomes. *Statistics in Medicine, 30*(8), 825–838.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). Praeger.

Chinni, M. L., & Hubley, A. M. (2014). A research synthesis of validation practices used to evaluate the Satisfaction with Life Scale (SWLS). In B. D. Zumbo & E. K. Chan (Eds.), Validity and validation in social, behavioral, and health sciences (pp. 35–66). Springer.

Christensen, A. P., Golino, H., & Silvia, P. J. (2020). A psychometric network perspective on the validity and validation of personality trait questionnaires. *European Journal of Personality, 34*(6), 1095–1108.

Cohen, J. (1990). Things i have learned (so far). *American Psychologist, 45*(12), 1304–1312.

Collie, R. J., & Zumbo, B. D. (2014). Validity evidence in the journal of educational psychology: Documenting current practice and a comparison with earlier practice. In B. D. Zumbo & E. K. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 113–135). Springer.

Coxe, S., & Sibley, M. H. (2023). Harmonizing DSM-IV and DSM-5 versions of ADHD "A Criteria": An item response theory analysis. *Assessment, 30*(3), 606–617.

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*(3), 242–247.
Curran, P. J., Cole, V. T., Bauer, D. J., Rothenberg, W. A., & Hussong, A. M. (2018). Recovering predictor-criterion relations using covariate-informed factor score estimates. *Structural Equation Modeling, 25*(6), 860–875.
Curran, P. J., McGinley, J. S., Bauer, D. J., Hussong, A. M., Burns, A., Chassin, L., & Zucker, R. (2014). A moderated nonlinear factor model for the development of commensurate measures in integrative data analysis. *Multivariate Behavioral Research, 49*(3), 214–231.
Curran, P. J., Cole, V., Bauer, D. J., Hussong, A. M., & Gottfredson, N. (2016). Improving factor score estimation through the use of observed background characteristics. *Structural Equation Modeling, 23*(6), 827–844.
DiStefano, C., Zhu, M., & Mindrila, D. (2019). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation, 14*(20), 1–11.
Donnellan, E., Usami, S., & Murayama, K. (2023). Random item slope regression: An alternative measurement model that accounts for both similarities and differences in association with individual items. *Psychological Methods*, advance online publication.
Edwards, M. C., & Wirth, R. J. (2009). Measurement and the study of change. *Research in Human Development, 6*(2–3), 74–96.
Edwards, K. D., & Soland, J. (2024). How scoring approaches impact estimates of growth in the presence of survey item ceiling effects. *Applied Psychological Measurement, 48*(3), 147–164.
Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher, 36*(8), 449–455.
Embretson, S. E. (2004). The second century of ability testing: Some predictions and speculations. *Measurement: Interdisciplinary Research and Perspectives, 2*(1), 1–32.
Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement, 20*(3), 201–212.
Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.
Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika, 82*, 904–927.
Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science, 16*(4), 779–788.
Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2015). COTAN review system for evaluating test quality. Retrieved February 19, 2024. https://www.psynip.nl/wp-content/uploads/2022/05/COTAN-review-system-for-evaluating-test-quality.pdf
Evers, A. (2012). The internationalization of test reviewing: Trends, differences, and results. *International Journal of Testing, 12*(2), 136–156.
Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure, and results. *International Journal of Testing, 10*(4), 295–317.
Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465.
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*(4), 370–378.
Flake, J. K. (2021). Strengthening the foundation of educational psychology by integrating construct validation into open science reform. *Educational Psychologist, 56*(2), 132–141.
Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *American Psychologist, 77*(4), 576–588.
Foster, G. C., Min, H., & Zickar, M. J. (2017). Review of item response theory practices in organizational research: Lessons learned and paths forward. *Organizational Research Methods, 20*(3), 465–486.
Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78*(2), 350.
Fried, E. I. (2020). Theories and models: What they are, what they are for, and what they are about. *Psychological Inquiry, 31*(4), 336–344.
Fried, E. I. (2015). Problematic assumptions have slowed down depression research: Why symptoms, not syndromes are the way forward. *Frontiers in Psychology, 6*, 309.
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), 1–11.
Fried, E. I., & Nesse, R. M. (2014). The impact of individual depressive symptoms on impairment of psychosocial functioning. *PloS One, 9*(2), e90311.
Gonzalez, O. (2021). Psychometric and machine learning approaches for diagnostic assessment and tests of individual classification. *Psychological Methods, 26*(2), 236–254.
Gonzalez, O., MacKinnon, D. P., & Muniz, F. B. (2021). Extrinsic convergent validity evidence to prevent jingle and jangle fallacies. *Multivariate Behavioral Research, 56*(1), 3–19.
Gorter, R., Fox, J. P., Riet, G. T., Heymans, M. W., & Twisk, J. W. R. (2020). Latent growth modeling of IRT versus CTT measured longitudinal latent variables. *Statistical Methods in Medical Research, 29*(4), 962–986.
Gorter, R., Fox, J. P., Apeldoorn, A., & Twisk, J. (2016). Measurement model choice influenced randomized controlled trial results. *Journal of Clinical Epidemiology, 79*, 140–149.

Gottfredson, N. C., Cole, V. T., Giordano, M. L., Bauer, D. J., Hussong, A. M., & Ennett, S. T. (2019). Simplifying the implementation of modern scale scoring methods with an automated R package: Automated moderated nonlinear factor analysis (aMNLFA). *Addictive Behaviors, 94*, 65–73.

Grice, J. W., & Harris, R. J. (1998). A comparison of regression and loading weights for the computation of factor scores. *Multivariate Behavioral Research, 33*(2), 221–247.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430–450.

Gunnell, K. E., Schellenberg, B. J., Wilson, P. M., Crocker, P. R., Mack, D. E., & Zumbo, B. D. (2014). A review of validity evidence presented in the journal of sport and exercise psychology (2002–2012): Misconceptions and recommendations for validation research. In B. D. Zumbo & E. K. Chan (Eds.), Validity and validation in social, behavioral, and health sciences (pp. 137-156). Springer.

Hair, J. F., Sharma, P. N., Sarstedt, M., Ringle, C. M., & Liengaard, B. D. (2024). The shortcomings of equal weights estimation and the composite equivalence index in PLS-SEM. *European Journal of Marketing, 58*(13), 30–55.

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). Scientific Software International.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika, 61*(4), 679–693.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62*(3), 331–347.

Higgins, W. C., Kaplan, D. M., Deschrijver, E., & Ross, R. M. (2023). Construct validity evidence reporting practices for the Reading the mind in the eyes test: A systematic scoping review. *Clinical Psychology Review, 108*, 102378.

Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64*(5), 802–812.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review, 14*(3), 332–346.

Howard, A. L. (2024). Graduate students need more quantitative methods support. *Nature Reviews Psychology, 3*, 140–141.

Hsiao, Y. Y., Kwok, O. M., & Lai, M. H. (2018). Evaluation of two methods for modeling measurement errors when testing interaction effects with observed composite scores. *Educational and Psychological Measurement, 78*(2), 181–202.

Huang, P. H. (2022). Penalized least squares for structural equation modeling with ordinal responses. *Multivariate Behavioral Research, 57*(2–3), 279–297.

Hubley, A. M., Zhu, S. M., Sasaki, A., & Gadermann, A. M. (2014). Synthesis of validation practices in two assessment journals: Psychological Assessment and the European Journal of Psychological Assessment. In B. D. Zumbo & E. K. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 193–213). Springer.

Hussong, A. M., Gottfredson, N. C., Bauer, D. J., Curran, P. J., Haroon, M., Chandler, R., & Springer, S. A. (2019). Approaches for creating comparable measures of alcohol use symptoms: Harmonization with eight studies of criminal justice populations. *Drug and Alcohol Dependence, 194*, 59–68.

Hwang, H., Cho, G., Jung, K., Falk, C. F., Flake, J. K., Jin, M. J., & Lee, S. H. (2021). An approach to structural equation modeling with both factors and components: Integrated generalized structured component analysis. *Psychological Methods, 26*(3), 273–294.

Jackson, D. L., Gillaspy, J. A., Jr., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods, 14*(1), 6–23.

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science, 15*(3), 809–816.

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling, 23*(4), 555–566.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.

Kang, S. M., & Waller, N. G. (2005). Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement, 29*(2), 87–105.

Kessels, R., Moerbeek, M., Bloemers, J., & van Der Heijden, P. G. (2021). A multilevel structural equation model for assessing a drug effect on a patient-reported outcome measure in on-demand medication data. *Biometrical Journal, 63*(8), 1652–1672.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43*(4), 355–381.

König, C., Khorramdel, L., Yamamoto, K., & Frey, A. (2021). The benefits of fixed item parameter calibration for parameter accuracy in small sample situations in large-scale assessments. *Educational Measurement: Issues and Practice, 40*(1), 17–27.

Kuhfeld, M., & Soland, J. (2022). Avoiding bias from sum scores in growth estimates: An examination of IRT-based approaches to scoring longitudinal survey responses. *Psychological Methods, 27*(2), 234–260.

Kuhfeld, M., & Soland, J. (2023). Scoring assessments in multisite randomized control trials: Examining the sensitivity of treatment effect estimates to measurement choices. *Psychological Methods*, advance online publication.

Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman–Brown to maximal reliability. *Psychological Methods, 1*(1), 98–107.

Li, X., & Jacobucci, R. (2022). Regularized structural equation modeling with stability selection. *Psychological Methods, 27*(4), 497–518.

Liang, X., & Jacobucci, R. (2020). Regularized structural equation modeling to detect measurement bias: Evaluation of lasso, adaptive lasso, and elastic net. *Structural Equation Modeling, 27*(5), 722–734.
Liu, Q., & Wang, L. (2021). t-Test and ANOVA for data with ceiling and/or floor effects. *Behavior Research Methods, 53*(1), 264–277.
Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584–585.
Luningham, J. M., McArtor, D. B., Bartels, M., Boomsma, D. I., & Lubke, G. H. (2017). Sum scores in twin growth curve models: Practicality versus bias. *Behavior Genetics, 47*, 516–536.
Maassen, E., D'Urso, E. D., van Assen, M. A., Nuijten, M. B., De Roover, K., & Wicherts, J. M. (2024). The dire disregard of measurement invariance testing in psychological science. *Psychological Methods*, advance online publication.
Maxwell, S. E., & Delaney, H. D. (1985). Measurement and statistics: An examination of construct validity. *Psychological Bulletin, 97*(1), 85–93.
McClure, K., Ammerman, B. A., & Jacobucci, R. (2024). On the selection of item scores or composite scores for clinical prediction. *Multivariate Behavioral Research, 59*(3), 566–583.
McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods, 52*, 2287–2305.
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, 55*(8), 4269–4290.
McNeish, D. (2023). Generalizability of dynamic fit index, equivalence testing, and Hu & Bentler cutoffs for evaluating fit in factor analysis. *Multivariate Behavioral Research, 58*(1), 195–219.
McNeish, D. (2023). Dynamic fit index cutoffs for categorical factor analysis with Likert-type, ordinal, or binary responses. *American Psychologist, 78*(9), 1061–1075.
Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the role of task model variables in assessment design. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 97–128). Lawrence Erlbaum.
Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), i–29.
Morgan-López, A. A., Saavedra, L. M., Hien, D. A., Norman, S. B., Fitzpatrick, S. S., Ye, A., & Back, S. E. (2023). Differential symptom weighting in estimating empirical thresholds for underlying PTSD severity: Toward a "platinum" standard for diagnosis? *International Journal of Methods in Psychiatric Research, 32*(3), e1963.
Morse, B. J., Johanson, G. A., & Griffeth, R. W. (2012). Using the graded response model to control spurious interactions in moderated multiple regression. *Applied Psychological Measurement, 36*(2), 122–146.
Müller, S., Hopwood, C. J., Skodol, A. E., Morey, L. C., Oltmanns, T. F., Benecke, C., & Zimmermann, J. (2023). Exploring the predictive validity of personality disorder criteria. *Personality Disorders: Theory, Research, and Treatment, 14*(3), 309–320.
Murray, A. L., Molenaar, D., Johnson, W., & Krueger, R. F. (2016). Dependence of gene-by-environment interactions (GxE) on scaling: Comparing the use of sum scores, transformed sum scores and IRT scores for the phenotype in tests of GxE. *Behavior Genetics, 46*, 552–572.
Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology, 73*, 719–748.
Padilla García, J. L., & Benítez Baena, I. (2014). Validity evidence based on response processes. *Psicothema, 26*(1), 136–144.
Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528–530.
Pelt, D. H., Schwabe, I., & Bartels, M. (2023). Bias in gene-by-environment interaction effects with sum scores: An application to well-being phenotypes. *Behavior Genetics, 53*, 359–373.
Peters, G. J., & Crutzen, R. (2024). Knowing what we're talking about: Facilitating decentralized, unequivocal publication of and reference to psychological construct definitions and instructions. *Meta-Psychology, 8*, 1–27.
Proust-Lima, C., Philipps, V., Dartigues, J. F., Bennett, D. A., Glymour, M. M., Jacqmin-Gadda, H., & Samieri, C. (2019). Are latent variable models preferable to composite score approaches when assessing risk factors of change? Evaluation of type-I error and statistical power in longitudinal cognitive studies. *Statistical Methods in Medical Research, 28*(7), 1942–1957.
Proust-Lima, C., Dartigues, J. F., & Jacqmin-Gadda, H. (2011). Misuse of the linear mixed model when evaluating risk factors of cognitive decline. *American Journal of Epidemiology, 174*(9), 1077–1088.
Pruzek, R. M., & Frederick, B. C. (1978). Weighting predictors in linear models: Alternatives to least squares and limitations of equal weights. *Psychological Bulletin, 85*(2), 254–266.
Qualls, A. L., & Moss, A. D. (1996). The degree of congruence between test standards and test documentation within journal publications. *Educational and Psychological Measurement, 56*(2), 209–214.
Ramsay, J. O., & Wiberg, M. (2017). A strategy for replacing sum scoring. *Journal of Educational and Behavioral Statistics, 42*(3), 282–307.
Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*(2), 93–103.
Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5*, 27–48.
Revelle, W. (2024). The seductive beauty of latent variable models: Or why I don't believe in the Easter Bunny. *Personality and Individual Differences, 221*, 112552.
Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods, 25*(1), 30–45.

Rodgers, J. L., & Shrout, P. E. (2018). Psychology's replication crisis as scientific opportunity: A précis for policymakers. *Policy Insights from the Behavioral and Brain Sciences, 5*(1), 134–141.
Rose, N., Wagner, W., Mayer, A., & Nagengast, B. (2019). Model-based manifest and latent composite scores in structural equation models. *Collabra: Psychology, 5*(1), 9.
Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in personality and social psychology bulletin. *Personality and Social Psychology bulletin, 28*(12), 1629–1646.
Schwabe, I., & van den Berg, S. M. (2014). Assessing genotype by environment interaction in case of heterogeneous measurement error. *Behavior Genetics, 44*(4), 394–406.
Schimmack, U. (2021). The validation crisis in psychology. *Meta-Psychology, 5*, 1–9.
Shaw, M., Cloos, L. J., Luong, R., Elbaz, S., & Flake, J. K. (2020). Measurement practices in large-scale replications: Insights from Many Labs 2. *Canadian Psychology/Psychologie Canadienne, 61*(4), 289.
Schreiber, J. B. (2021). Issues and recommendations for exploratory factor analysis and principal component analysis. *Research in Social and Administrative Pharmacy, 17*(5), 1004–1011.
Shear, B. R., & Zumbo, B. D. (2014). What counts as evidence: A review of validity studies in educational and psychological measurement. In B. D. Zumbo & E. K. Chan (Eds.), Validity and validation in social, behavioral, and health sciences (pp. 91–111). Springer.
Sijtsma, K. (2012). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika, 77*, 4–20.
Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107–120.
Sijtsma, K., Ellis, J. L., & Borsboom, D. (2024). Recognize the value of the sum score, psychometrics' greatest accomplishment. *Psychometrika, 89*(1), 84–117.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.
Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26*(1), 100–107.
Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*, 83–117.
Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5*(4), 299–321.
Slof-Op't Landt, M. C. T., van Furth, E. F., Rebollo-Mesa, I., Bartels, M., van Beijsterveldt, C. E. M., Slagboom, P. E., & Dolan, C. V. (2009). Sex differences in sum scores may be hard to interpret: The importance of measurement invariance. *Assessment, 16*(4), 415–423.
Soland, J. (2022). Evidence that selecting an appropriate item response theory-based approach to scoring surveys can help avoid biased treatment effect estimates. *Educational and Psychological Measurement, 82*(2), 376–403.
Soland, J., Kuhfeld, M., & Edwards, K. (2022a). How survey scoring decisions can influence your study's results: A trip through the IRT looking glass. *Psychological Methods*, advance online publication.
Soland, J., McGinty, A., Gray, A., Solari, E. J., Herring, W., & Xu, R. (2022). Early literacy, equity, and test score comparability during the pandemic. *Educational Assessment, 27*(2), 98–114.
Soland, J., Johnson, A., & Talbert, E. (2023). Regression discontinuity designs in a latent variable framework. *Psychological Methods, 28*(3), 691–704.
Soland, J., Cole, V., Tavares, S., & Zhang, Q. (2024). Evidence that growth mixture model results are highly sensitive to scoring decisions. *PsyArXiv*. https://osf.io/preprints/psyarxiv/d27rcSpeelman
Speelman, C. P., Parker, L., Rapley, B. J., & McGann, M. (2024). Most psychological researchers assume their samples are ergodic: Evidence from a year of articles in three major journals. *Collabra: Psychology, 10*(1), 92888.
Stochl, J., Fried, E. I., Fritz, J., Croudace, T. J., Russo, D. A., Knight, C., & Perez, J. (2022). On dimensionality, measurement invariance, and suitability of sum scores for the PHQ-9 and the GAD-7. *Assessment, 29*(3), 355–366.
Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science. *Annual Review of Clinical Psychology, 15*, 579–604.
Tang, X., Schalet, B. D., Peipert, J. D., & Cella, D. (2023). Does scoring method impact estimation of significant individual changes assessed by patient-reported outcome measures? Comparing classical test theory versus item response theory. *Value in Health, 23*(10), 1518–1524.
Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining. *European Journal of Personality, 34*(5), 826–844.
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.
Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*(2), 371–390.
van den Oord, E. J., Pickles, A., & Waldman, I. D. (2003). Normal variation and abnormality: An empirical study of the liability distributions underlying depression and delinquency. *Journal of Child Psychology and Psychiatry, 44*(2), 180–192.
van den Oord, E. J., & van der Ark, L. A. (1997). A note on the use of the Tobit approach for tests scores with floor or ceiling effects. *British Journal of Mathematical and Statistical Psychology, 50*(2), 351–364.
van der Ark, L. A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika, 70*, 283–304.
Vogelsmeier, L. V., Jongerling, J., & Maassen, E. (2024). Assessing and accounting for measurement in intensive longitudinal studies: Current practices, considerations, and avenues for improvement. *Quality of Life Research*, advance online publication.
Vogelsmeier, L. V., Vermunt, J. K., Keijsers, L., & De Roover, K. (2021). Latent Markov latent trait analysis for exploring measurement model changes in intensive longitudinal data. *Evaluation & the Health Professions, 44*(1), 61–76.

Vogelsmeier, L. V., Vermunt, J. K., van Roekel, E., & De Roover, K. (2019). Latent Markov factor analysis for exploring measurement model changes in time-intensive longitudinal studies. *Structural Equation Modeling, 26*(4), 557–575.
Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83*(2), 213–217.
Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion, 17*(2), 267–295.
Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R., & van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7*, 1832.
Wilson, M., Allen, D. D., & Li, J. C. (2006). Improving measurement in health education and health behavior research using item response modeling: comparison with the classical test theory approach. *Health Education Research, 21*(supplement 1), i19–i32.
Wolf, M. G. (2023). The problem with over-relying on quantitative evidence of validity. *PsyArXiv*. https://doi.org/10.31234/osf.io/v4nb2
Zwitser, R. J., & Maris, G. (2016). Ordering individuals with sum scores: The introduction of the nonparametric Rasch model. *Psychometrika, 81*, 39–59.

Manuscript Received: 1 APR 2024
Published Online Date: 20 JUL 2024