Evaluating When Subscores Can Have Value in Psychological and Health Applications

by

Molly Gardner

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Approved January 2022 by the
Graduate Supervisory Committee:

Michael C. Edwards, Chair
Daniel McNeish
Roy Levy

ARIZONA STATE UNIVERSITY

May 2022

ABSTRACT

Scale scores play a significant role in research and practice in a wide range of areas such as education, psychology, and health sciences. Although the methods of scale scoring have advanced considerably over the last 100 years, researchers and practitioners have generally been slow to implement these advances. There are many topics that fall under this umbrella but the current study focuses on two. The first topic is that of subscores and total scores. Many of the scales in psychological and health research are designed to yield subscores, yet it is common to see total scores reported instead. Simplifying scores in this way, however, may have important implications for researchers and scale users in terms of interpretation and use. The second topic is subscore augmentation. That is, if there are subscores, how much value is there in using a subscore augmentation method? Most people using psychological assessments are unfamiliar with score augmentation techniques and the potential benefits they may have over the traditional sum score approach. The current study borrows methods from education to explore the magnitude of improvement of using augmented scores over observed scores. Data was simulated using the Graded Response Model. Factors controlled in the simulation were number of subscales, number of items per subscale, level of correlation between subscales, and sample size. Four estimates of the true subscore were considered (raw, subscore-adjusted, total score-adjusted, joint score-adjusted). Results from the simulation suggest that the score adjusted with total score information may perform poorly when the level of inter-subscore correlation is 0.3. Joint scores perform well most of the time, and the subscore-adjusted scores and joint-adjusted scores were always better performers than raw scores. Finally, general advice to applied users is provided.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Scale scores play a significant role in research and practice in a wide range of areas such as education, psychology, and health sciences. Although the methods of scale scoring have advanced considerably over the last 100 years, researchers and practitioners have generally been slow to implement these advances. There are many topics that fall under this umbrella, but we will focus on two in the current study.

The first topic is that of subscores and total scores. Many of the scales in psychological and health research are designed to yield subscores, yet we often see total scores reported instead. Simplifying scores in this way, however, may have important implications for researchers and scale users in terms of interpretation and use. That is, are there conditions where it is appropriate to use total scores instead of subscores and if so, how do those impact the subsequent interpretation and use of the scale score? Haberman (2008) and Sinharay (2010) have considered this from what amounts to the opposite perspective, i.e. under what conditions can subscores have value? Their work in the educational sector revealed that, in general, subscores can be valuable when they are reliable (i.e., consist of at least 20 items) and the (disattenuated) correlation among subscores is low (i.e., $r < 0.85$). We aim to replicate these findings, but also to extend the conditions studied by previous authors to those that more closely reflect what we see in psychology and health assessment.

The second topic is subscore augmentation. That is, if there are subscores, how much value is there in using a subscore augmentation method? Most people using psychological assessments are unfamiliar with score augmentation techniques and the potential benefits they may have over the traditional summed-score approach. Therefore, we will start with the simplest form of augmentation: Kelley's regressed estimate (Kelley, 1923) which shrinks scores to the sample mean proportional to their unreliability. The terms “raw” and “observed” will be used interchangeably throughout this paper to refer to the summed-score. The terms “augmented” and “adjusted” will be used interchangeably to refer to any score that has been changed from its observed form via Kelley’s formula or variants thereof. We are interested in exploring the magnitude of improvement using these types of augmented scores over observed scores, as well as whether the circumstances under which subscores add value are different for augmented versus observed scores.

Our goals for the current study are to: 1) conduct a simulation that enables me to make general claims about where subscores are more (or less) likely to be valuable in psychology and health research, 2) provide researchers with a functional tool that helps them assess the cost of simplifying their scores (e.g., subscore vs. total score; augmented vs. non-augmented), and 3) spark a larger conversation about scoring practices and techniques. In the next section of this document, we provide a brief review of classical test theory (CTT, Lord & Novick, 1968) and discuss the existing literature pertaining to the added value of subscores. After this, we describe the simulation and analytic strategy, and provide results. We conclude with a discussion of the results and offer general advice to applied users.

Review of Classical Test Theory & Reliability

The foundation of CTT (Lord & Novick, 2008) rests on the true score model, which states that an observed score x_j for individual j is the sum of two components: a true score τ_j and error e_j . This can be written as $x_j = \tau_j + e_j$. The true score τ_j is defined as the value we would expect to obtain if a person took parallel forms of an assessment infinitely many times and those scores were averaged (Lord & Novick, 2008; Wainer & Thissen, 2009, Chapter 2). It can also be thought of as an individual’s true level of the construct of interest. Error is defined to be random and characterized as the difference between the observed score x_j and the true score τ_j . The true score model assumes that:

(1) $E(x_j) = \tau_j$, the expected value (taken over hypothetical parallel replications) of an examinee's observed score is equal to their true score (the observed score is an unbiased estimate of the true score)

(2) $E(e_j) = 0$, the expected value of error scores is zero

(3) $\sigma_{\tau e} = 0$, the covariance between the true score and error is zero.

In general, *reliability* refers to how consistently an observed score on a particular measure captures the construct of interest (i.e., true score). The smaller the difference between observed score and true score (i.e., the smaller the error), the more reliable the observed score is. Formally, reliability is defined as the proportion of observed score variance that is true score variance, which is equal to the proportion that is *not* error variance, which is equal to the squared correlation between observed score and true score, $\rho_{x\tau}^2 = \frac{\sigma_{x\tau}^2}{\sigma_x^2 \sigma_\tau^2}$ (Wainer & Thissen, 2009, Chapter 2). However, because the true score is unobservable, it follows that reliability cannot be directly computed. Therefore, it must be estimated. The most widely used estimate of reliability is coefficient alpha (henceforth reliability, or α; Hoyt, 1941; Guttman, 1945; Cronbach, 1951), which assesses the internal consistency among observed item responses. For a k-item scale that has a unit-weighted sum score called *x,* coefficient α for the observed responses is expressed as,

$$\alpha = \frac{k}{k-1} \, (1 - \frac{\sum \sigma_k^2}{\sigma_x^2})$$

where $\sum \sigma_k^2$ is the sum of the item variances, σ_x^2 is the variance of the scale scores (which is equal to the sum of all the item variances and covariances), and $\frac{k}{k-1}$ restricts the estimate to be between 0 and 1. Thus, the formula for coefficient α is equal to the proportion of a scale's total variance that is attributable to a common source, assumed to be the true score. The higher α is, the more reliable the score is, such that if $\alpha = 1$, then the observed score is driven entirely by the true score, and if $\alpha = 0$, then none of the true score is being reflected in the observed score. A is considered an unbiased estimate of internal consistency if all the items comprising the scale are at least tau-equivalent. Tau equivalency requires that each item of a given scale measure the construct equally well (Novick & Lewis, 1967). If they are not all at least tau equivalent, then coefficient α has been shown to be a lower-bound estimate of reliability (Guttman, 1945; Miller, 1995). Additionally, coefficient α assumes that the scale is unidimensional (measures one construct), items are normally distributed, and the errors of the items do not covary (see McNeish, 2018).

In addition to estimating the reliability for a set of scores, the true scores also must be estimated, and there are various ways of doing this. The predominant approach in psychological and health research is to use the observed summed score, which is a sum of the scored item responses. It is also common to see researchers use an observed average as the scale score, which is the average over the observed item responses. We will focus on the summed score for this project, but as the average score is a monotonic transformation of the summed score, so the same conclusions should hold. It is worth acknowledging, however, that the monotonic transformation of the summed score to a mean score can break down with missing data. A total score (x_T) can be computed by summing all the item responses. Subscores (x_S) can be created by summing specific sets of item responses. For every assessment, there is one total score, x_T, and as many subscores, x_S, as there are subscales. Furthermore, any of the observed subscores can be subtracted from the observed total score to yield a difference that is referred to as the *observed remainder* score[1], $x_R = x_T - x_S$. There are as many observed remainder scores as there are subscores. In the CTT summed score approach, the observed total score x_T, observed subscores x_S, and observed remainder scores x_R are used as estimates of the true total score τ_T, true subscores τ_S, and true remainder scores τ_R, respectively.

Haberman's Method & Motivation

Psychological and health research are not the only sectors affected by the issue of total score versus subscores. A similar, yet distinct, issue has been observed and addressed in the educational context. Many standardized educational tests are designed to produce a single total score that is thought to represent a student's ability in that domain. However, since the No Child Left Behind Act of 2001 and Every Student Succeeds Act of 2015, there has been pressure to report subscores in addition to the total score on standardized tests (e.g., Feinberg & Jurich, 2017; Haberman, 2008; Sinharay, 2010). This pressure was driven by the notion that subscores may provide remedial and instructional benefits. To be of any diagnostic value however, subscores need to meet certain standards. Namely, Standards 1.13 and 1.14 of the *Standards for Educational and Psychological Testing* require that:

> When a test provides more than one score, the distinctiveness and reliability of the separate scores should be demonstrated, and the interrelationships of those scores should be shown to be consistent with the construct(s) being assessed. When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. (AERA et al., 2014, p. 27)

[1] While the remainder score may not be a particularly useful or familiar score to most researchers, it serves an important role when determining the cost of using a total score over subscores; this is discussed further in the Value-Added Ratio section.

To determine whether subscores should be reported, Haberman (2008) proposed a method that is grounded in CTT and is described next.

From the CTT perspective, the observed subscore is treated as an estimate of the true subscore. When determining whether subscores are worth reporting, Haberman (2008) considered three estimates of the true subscore:

1) $A_S = \bar{x}_S + \alpha_S(x_S - \bar{x}_S),$ which is a Kelley regressed estimate that is based on the observed subscore (x_S), the average subscore for the sample $(\bar{x}_S)$, and the reliability of the subscore (α_s).
2) $A_T = \bar{x}_S + c(x_T - \bar{x}_T)$, which is based on the observed total score (x_T), the average total score for the sample $(\bar{x}_T)$, and a constant (c) that is determined by the reliabilities and standard deviations of the subscore and total score and the correlations between the subscores.
3) $A_{TS} = \bar{x}_S + \alpha_S(x_S - \bar{x}_S) + b(x_T - \bar{x}_T)$, which is based on a weighted average of the observed subscore (x_S), the observed total score (x_T), the average subscore for the sample $(\bar{x}_S)$, the average total score for the sample $(\bar{x}_T)$, and two constants (a) and (b) that depend on the reliabilities and standard deviations of the subscore and total score and the correlations between the subscores.

Haberman (2008) used the mean-squared error (MSE) to evaluate the various scores. The MSE indicates the average squared distance between the estimated subscore and the true subscore. Because the true subscore is unknown, Haberman devised methods for calculating the quantities needed to estimate the necessary MSEs. The smaller the MSE, the more accurate the corresponding estimate is. In Haberman's case, the estimate was any one of the three augmented scores listed above. The MSE of each estimate was compared to the MSE of a baseline model, which Haberman defined as the mean of the observed subscores. This is equivalent to an intercept-only regression model. Rather than just relying on MSE, Haberman took additional steps to create a more interpretable summary of the performance of a given score; he did so by evaluating the *proportional reduction in MSE* (PRMSE) of the approximation of the true subscore by any one of the three augmented scores relative to the approximation by the sample mean. Thus, the PRMSE is the relative decrease in MSE from using any one of the three augmented scores compared to using the sample mean for all individuals:

$$PRMSE_A = 1 - \frac{MSE_A}{MSE_{E(\bar{x}_s)}} = \frac{MSE_{E(\bar{x}_s)} - MSE_A}{MSE_{E(\bar{x}_s)}}$$

There is a PRMSE value associated with each of the three estimates mentioned above: $PRMSE_{A_S}$, $PRMSE_{A_T}$, and $PRMSE_{A_{TS}}$, respectively. $PRMSE_{A_S}$ has been shown to be equal to the estimated reliability of the subscore (Haberman, 2008). The PRMSE values are bounded between 0 and 1, such that the larger the PRMSE value is, the more accurate the corresponding estimate is (Haberman, 2008; Sinharay, 2010; Sinharay et al., 2011b). For subscores to have value above and beyond that of the total score, Haberman determined that $PRMSE_{A_S}$ needs to be greater than $PRMSE_{A_T}$. Said differently, A_S needs to be a more accurate estimate of the true subscore than A_T. In sum, Haberman (2008) found that subscores are "most likely to have value if they have relatively high reliability by themselves and if the true subscore and true total score have only a moderate correlation. Both conditions are important" (p. 224).

Sinharay's Simulation

To quantify the findings of Haberman (2008)—specifically, how reliable and distinct subscores must be—Sinharay (2010) looked at operational and simulated datasets. Here, we will focus only on the simulation study that Sinharay conducted rather than the operational analyses, because that is most relevant to the current study. Sinharay (2010) used the same three estimates of the true subscore that Haberman (2008) did: one that is based on the observed subscore (A_S), another that is based on the observed total score (A_T), and a third based on a weighted average of the observed subscore and observed total score (A_{TS}). This weighted average "places the same weight on the subscores other than the one of interest" and is a special case of Wainer et al's (2009) augmented subscore, which "places different weights on all the subscores" (Sinharay, 2010, p.152). Sinharay (2010) used the 2-parameter logistic multidimensional item response theory model (2-PL MIRT model; Reckase, 2007) to generate dichotomous item response data. Then he applied Haberman's PRMSE method to evaluate the performance of the three augmented scores in estimating the true subscore. The factors and levels of each factor that Sinharay controlled in his simulation are as follows:

- Number of subscales: 2, 3, and 4
- Length of subscales: 10, 20, 30, and 50
- Level of correlation among subscales: .70, .75, .80, .85, .90, and .95
- Sample size N: 100, 1,000, and 4,000

Results from the simulation study suggested that the biggest factors impacting the value of subscores are, as expected, the average reliability and average disattenuated correlation among subscores. According to Sinharay's (2010) results, subscores need to consist of at least 20 items to be considered reliable. For example, for 2, 3, and 4 subscales with 10 items each, the average reliability was 0.58; the subscores were not of any added value even when they were sufficiently distinct from one another (i.e., disattenuated inter-subscore correlation of 0.7). But for subscores with 20 items each, the average reliability was 0.74; for 30 items each the average reliability was 0.81, and for subscores with 50 items each, the average reliability was 0.88. Furthermore, for subscores to be considered sufficiently distinct from each other, the disattenuated correlations among subscores should be less than 0.85. Sinharay also pointed out that "there is an interaction between the length of the subscores and the level of correlation" (p. 167). For example, the added value of subscales that consist of at least 20 items depends on the level of correlation among the subscores. These results reflect educational conditions, however, which is the motivation for the current study. That is, will the results be the same if the conditions are manipulated to reflect what we see in psychological and health research?

The Value-Added Ratio

Building on the work of Haberman (2008) and Sinharay (2010), Feinberg and Wainer (2014) proposed examining the ratio of $PRMSE_{A_S}$ to $PRMSE_{A_T}$ and called this the value-added ratio (VAR). Although VAR could be calculated as the ratio between estimates of the two necessary PRMSE values, Feinberg and Wainer also derived a simpler equation to estimate the VAR directly. The equation, shown below, was derived from the results of Feinberg (2012):

$$\frac{PRMSE_{A_S}}{PRMSE_{A_T}} = VAR \approx 1.15 + 0.5 \times r_1 - 0.67 \times r_2,$$

where r_1 is the reliability of the subscore and r_2 is the disattenuated correlation of the subscore with the remainder score. In other words, r_2 is the raw correlation between the subscore and the remainder of the test divided by the square root of the product of their reliabilities (Feinberg & Wainer, 2014; Feinberg & Jurich, 2017). If VAR is greater than one, then A_S is more valuable than A_T because it is explaining more variance in the true subscore relative to A_T. If VAR is less than one, then A_S is less valuable than A_T in predicting the true subscore (Feinberg & Jurich, 2017; Feinberg & Wainer, 2014). To evaluate and interpret the magnitude of VAR, Feinberg & Jurich (2017) tested its statistical and practical significance. A ratio test (Wilcox & Tian, 2008) was used to examine the statistical significance of VAR and after transforming PRMSE values to z-scores, compared to Cohen's q criteria for an effect size estimate. According to Feinberg and Jurich (2017) a VAR value greater than one is typically necessary for a subscore to add value. Further, they go on to suggest that scores with a VAR <0.9 may be harmful.

The VAR may not be a natural metric for researchers to understand the relative value of using (or not using) subscores. There is also concern that the equation used to estimate VAR might not be accurate (Sinharay et al., 2015). As part of this study, we intend to evaluate the performance of the VAR simplification as well as consider the utility of VAR in communicating with applied researchers.

Validity & Dimensionality

The degree to which scales are multidimensional has been a long-standing debate amongst applied researchers and psychometricians. Many scales—particularly in psychological and health research—are designed to yield subscores. That is, the constructs are conceptualized as being comprised of multiple facets. There is an entire literature on how to determine dimensionality and work has been done to support the use of subscores on multidimensional scales. Therefore, the current study adopts the perspective that there is existing evidence of multidimensionality and the methods by which that is determined are not discussed here. The current study is looking at multidimensional scales that use summed scoring and there has been debate (and/or stark inconsistencies across the studies using that scale) about whether to report subscores or a total score.

Furthermore, the augmentation techniques to increase subscore precision have been criticized for lacking validity (e.g., Skorupski & Carvajal, 2010; Stone et al., 2010). There has been concern that once a subscore has been augmented, it no longer measures the construct of interest in a valid way. These concerns have been addressed by Sinharay et al. (2011a), who showed that subscores do not lose their meaning once adjusted, nor do they misrepresent the construct being measured. These are interesting and worthwhile arguments to consider, and will be addressed briefly in the discussion section.

CHAPTER 2

METHODS

The Simulating Model

We will use the Graded Response Model (GRM; Samejima, 1969) to generate item response data for items with five response categories. Polytomous items that employ an ordered response scale are pervasive in psychological research. The GRM is appropriate for items with ordered response options and expresses the probability of responding in a particular category. For an item with m response options, the probability of responding in the c^{th} category is expressed as,

$$P(x_i = c|\,\theta) = \frac{1}{1 + exp[-a_i(\theta - b_{ic})]} - \frac{1}{1 + \exp[-a_i(\theta - b_{ic+1})]}\ ,$$

where x_i is the observed response to item i, a_i is the slope parameter for item i, b_i is the severity parameter (also known as difficulty in the education context) for item i, and θ is the construct being measured. The slope parameter, a_i, represents the degree to which the item is related to θ; the severity parameter, b_{ic}, is the level of θ required for an individual to have a 50% chance of endorsing response option c or higher for item i. Each item has one a-parameter and m-1 b-parameters associated with it. For this study, m is set to five, so there are four b-parameters for each polytomous item.

Generating Parameters

Theta values were generated from a multivariate Normal distribution with mean vector $\boldsymbol{\mu}$ = **0** and covariance matrix $\boldsymbol{\Sigma}$.We set the diagonals of $\boldsymbol{\Sigma}$ to one and the off diagonals were the correlations between the dimensions of theta. The a-parameter was sampled from a Normal distribution with a mean of 1.7 and standard deviation of 0.3. This choice was based on the a-parameter distribution found by Hill's (2004) review of 15 published articles which applied the GRM to psychological scales. The four needed b-parameters (m-1) were constructed according to the procedure outlined in Hill (2004).

That is, the first *b*-parameter was drawn from a Normal distribution with a mean of -1.5 and a standard deviation of 0.5. Then, to obtain the remaining *b*-parameters, "shift" values (restricted to positive numbers) were drawn from a *N*(1.0, 0.2) distribution and added to the previous *b*-parameter value. For example, if the first *b*-parameter is drawn as -1.3 and the first shift value is 0.7, then the second *b*-parameter would be (-1.3 + 0.7 = -0.6). A second random-draw shift value of 0.2 would yield a third *b*-parameter of (-0.6 + 0.2 = -0.4), and a third random-draw shift value of 0.9 would yield a fourth *b*-parameter equal to (-0.4 + 0.9 = 0.5). As a reminder, Sinharay (2010) did not need to use shift values because he generated data from the 2-PL model, which only requires one *b*-parameter per item.

The true scores are calculated for each simulee based on the simulated items they interacted with. This is the same procedure used to generate what is often called an expected response function, which links the theta metric to the observed score metric. For each item, the expected observed response is the sum of the products of the probability of choosing a category and the category value itself. For each subscale, the item-level expectations are summed, which provides an expected subscore. This is the expected summed score given the item parameters and theta value for the simulee in question and this value is used as each simulee's true score.

Factors Controlled in the Simulation

The factors varied in this simulation mimic those of Sinharay (2010). However, the levels of these factors are slightly different than those used by Sinharay (2010) because they are adjusted to mimic what is commonly seen in psychological and health outcome research, rather than in education. For example, Sinharay did not consider observed scores as estimates of the true subscore. Therefore, we expect to learn about the performance of observed subscores compared to augmented subscores. We controlled the following factors in the simulation:

- Number of subscales (i.e., dimensions): 2 and 5
- Length of the subscores: 4 and 8 items per subscale
- Level of latent correlations among subscores: 0.3, 0.7, and 0.9
- Sample size *N*: 250
- Type of score: augmented and non-augmented (i.e., raw/observed)

The rationale for choosing these factors and levels was guided by a review of the literature. There was a total of 12 conditions simulated and 100 replications per cell (condition).

Analytic Strategy

When examining the data, we considered four estimates of the desired subscore. We used the three scores studied by Haberman (2008) and Sinharay (2010) as well as included the observed subscore (x_S). The three scores examined in previous work were included on their own merits as well as to enable easy comparison to previous findings. The raw score was added to keep the simulation more faithful to what is commonly found in psychology and health outcomes. In these fields, it is incredibly rare to see any kind of weighted score along the lines of a Kelley regressed estimate. This could be due to general unfamiliarity with weighted scoring methods and the potential benefits they have over the summed-score approach. Or it could be that people are aware of weighted scores (and their benefits) but lack experience in calculating them and are therefore deterred away from adopting these methods.

To examine the quality of subscore estimates, we split the analyses by number of items per subscale. Although splitting by number of items limits the ability to detect the specific effect of reliability on RMSE (because number of items is a proxy for reliability), it does enable us to see how the number of dimensions, score type, and level of correlation among subscores affects the RMSE across shorter and longer subscales. That is, we are not considering the specific impact of reliability in the current study, but rather looking at two subscale lengths using item parameters with empirical support. For each analysis, we conducted a factorial ANOVA with the RMSE of subscore estimates as the outcome variable and factors corresponding to number of dimensions, level of correlation among subscores, and score type. All possible interactions were included in the ANOVA model. ANOVA model assumptions (independence of observations, normal distributions, and homogeneity of variance) were visually examined and no obvious violations were observed. To assess the variance accounted for by each factor, semipartial eta-squared ($\eta^2_{partial}$) was used as the measure of effect size, which is a relative measure of variance accounted for. That is, relative to the total amount of variance accounted for by the model, how much was due to each factor.

In order to examine the conditions under which subscores have added value over the total score and to assess if that added value depends on whether the scores are augmented, we modified Haberman's PRMSEs. The PRMSE, by its nature, provides an estimate of the relative improvement in MSE by considering the estimate in the numerator rather than the estimate in the denominator. We modified Haberman's PRMSEs such that the subscore sample mean was no longer used as the baseline/reference to which the other MSEs are being compared. Instead, the MSE associated with the total score and the MSE associated with the observed subscore were used in the denominator for comparisons. Each MSE comparison (henceforth, MSEC) is described next.

In order to make general claims about when subscores are more (or less) likely to have added value relative to the total score, we examined $MSEC_{S/_{T}}$. The larger this value is, the greater the reduction in MSE is from using the observed subscore, proportional to using the total score as an estimate of the true subscore. By using the MSE associated with the total score instead of the MSE associated with the subscore sample mean in the denominator, we can directly compare the raw subscore to the total score. $MSEC_{AS/_S}$, $MSEC_{AT/_S}$, and $MSEC_{ATS/_S}$ enabled us to examine the added value of using adjusted subscores (of any sort) over observed subscores. As with the first modification, these modifications facilitate meaningful and relevant comparisons. Additionally, $MSEC_{AS/_T}$, $MSEC_{AT/_T}$, and $MSEC_{ATS/_T}$ can help us understand how augmentation of various types impacts the value of the subscore relative to the total score.

We hypothesized that the raw subscores will always be a more accurate estimate of the true subscore than the total score because the raw subscore is not obscured by the additional information that is contained within the total score. We expect that the adjusted subscores will always outperform the observed subscores in terms of MSE because they are incorporating relevant information, which should improve prediction. Furthermore, out of the three augmented scores, we expect the ATS score to always outperform the AS and AT scores because it is taking into account the most information. By making all of these comparisons, we hope to provide researchers with a functional tool that helps them assess the cost of simplifying their scores (e.g., subscore vs. total score, adjusted vs. observed) as well as spark a larger conversation about scoring practices and techniques.

Lastly, the various adjusted scores depend in part on the estimates of reliability and inter-subscore correlations. We examined both RMSE and bias for the reliability and inter-subscore correlation estimates. The RMSE of raw and disattenuated inter-subscore correlation estimates were computed by subtracting the generating correlation (which was either 0.3, 0.7, or 0.9 depending on the condition) from the corresponding estimated correlation. This value was then squared and then all the squared values were averaged, resulting in the MSE. Finally, taking the square-root of the MSE yielded the RMSEs of the inter-subscore correlation estimates.

CHAPTER 3

RESULTS

RMSE of Subscore Estimates

Tables 1 through 4 display the results for the RMSE of subscore estimates, bolded entries in all tables indicate that the *F*-test for that effect was significant at an α level of 0.05. Based on Cohen (1988), eta-squared values of 0.01, 0.06, and 0.14 correspond to small, medium, and large effects, respectively.

For subscales with four items, the overall model accounted for 94% of variance in RMSE of subscore estimates, $\eta^2 = 0.94, CI[(0.93, 0.94)]$. Table 1 indicates that, of the total explained variance, 22% was due to level of inter-subscore correlation; 33% was due to score type; and 35% was due to the interaction between level of inter-subscore correlation and score type. The sum of these three larger effects is 0.90.

Table 1

ANOVA Results for Each Manipulated Factor when K=4

Source	*df*	Semipartial η²	95% CI	
			Lower	Upper
D	1	0.00	0.00	0.00
R	2	0.22	0.19	0.25
Score Type	3	0.33	0.30	0.36
D × R	2	0.01	0.00	0.02
D × Score Type	3	0.01	0.00	0.02
R × Score Type	6	0.35	0.32	0.37
D × R × Score Type	6	0.02	0.01	0.03

K = Items per Subscale, D = Dimension, R = Level of Inter-subscore Correlation

For subscales with eight items, the overall model accounted for 98% of variance in RMSE of subscore estimates, $\eta^2 = 0.98, CI[(0.98, 0.98)]$. Table 2 indicates that, of the total explained variance, 19% was due to level of inter-subscore correlation; 39% was due to score type; and 36% was due to the interaction between level of inter-subscore correlation and score type. The sum of these three effects is 0.94.

Table 2
ANOVA Results for Each Manipulated Factor when K=8

Source	*df*	Semipartial η²	95% CI	
			Lower	Upper
D	1	0.00	0.00	0.01
R	2	0.19	0.16	0.22
Score Type	3	0.39	0.37	0.42
D × R	2	0.01	0.00	0.01
D × Score Type	3	0.01	0.01	0.02
R × Score Type	6	0.36	0.33	0.39
D × R × Score Type	6	0.01	0.00	0.02

K = Items per Subscale, D = Dimension, R = Level of Inter-subscore Correlation

In both cases (K=4 and K=8), the Dimension factor accounted for less than 0.5% of the total explained variance, indicating that, at least in this simulation, the number of dimensions has little impact on the RMSE of the subscore estimate. Therefore, the remainder of the results reported for subscore estimates will aggregate over the Dimension factor.

Averaging over the Dimension factor for four-item subscales, Table 3 and Figure 1 show the RMSE Least-Squares Mean (aka LS means, marginal means, estimated marginal means) for each score type by level of inter-subscore correlation effect (12 effects total).

Table 3
Average RMSE LS Mean for Subscore Estimates when K=4

Score Type	Level of Correlation		
	0.3	0.7	0.9
AS	1.64	1.64	1.64
AT	2.41	1.78	1.34
ATS	1.61	1.48	1.29

Raw	1.9	1.89	1.9

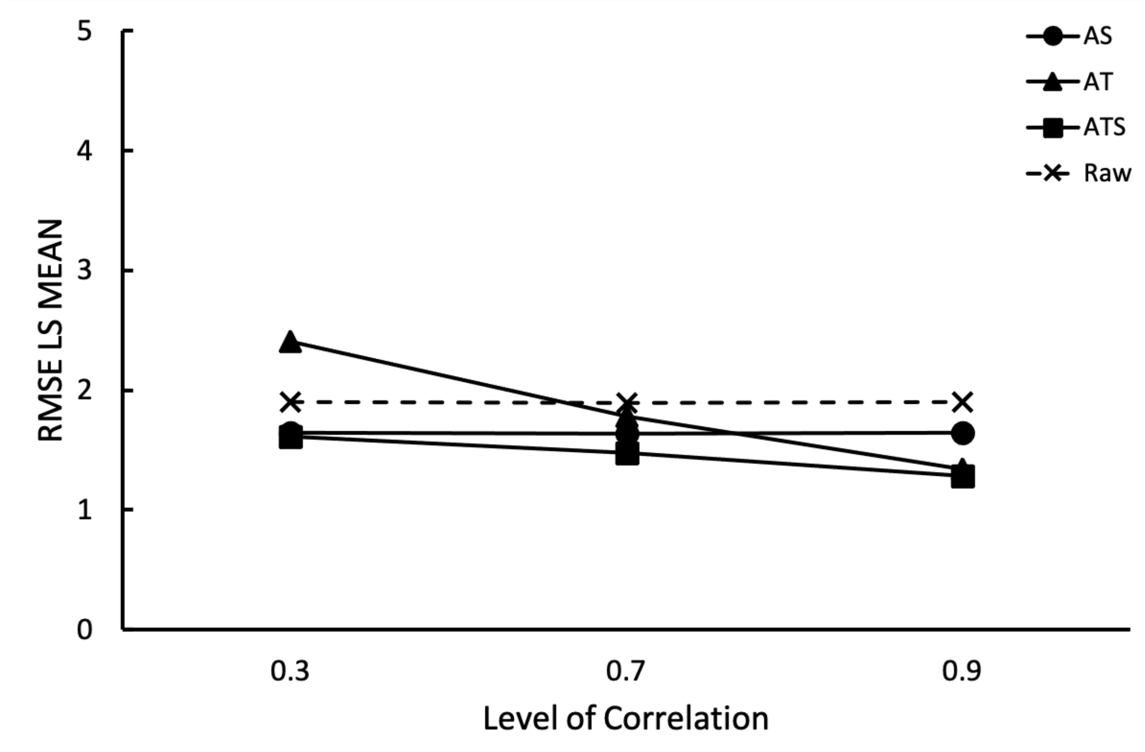

Figure 1. Average RMSE LS Mean for Subscore Estimates when K=4

Averaging over the dimension factor for eight-item subscales, Table 4 and Figure 2 display the RMSE LS Mean for each score type by level of inter-subscore correlation effect (12 effects total).

Table 4
Average RMSE LS Mean for Subscore Estimates when K=8

Score Type	Level of Correlation		
	0.3	0.7	0.9
AS	2.49	2.48	2.47
AT	4.61	3.24	2.26
ATS	2.46	2.31	2.02
Raw	2.69	2.68	2.68

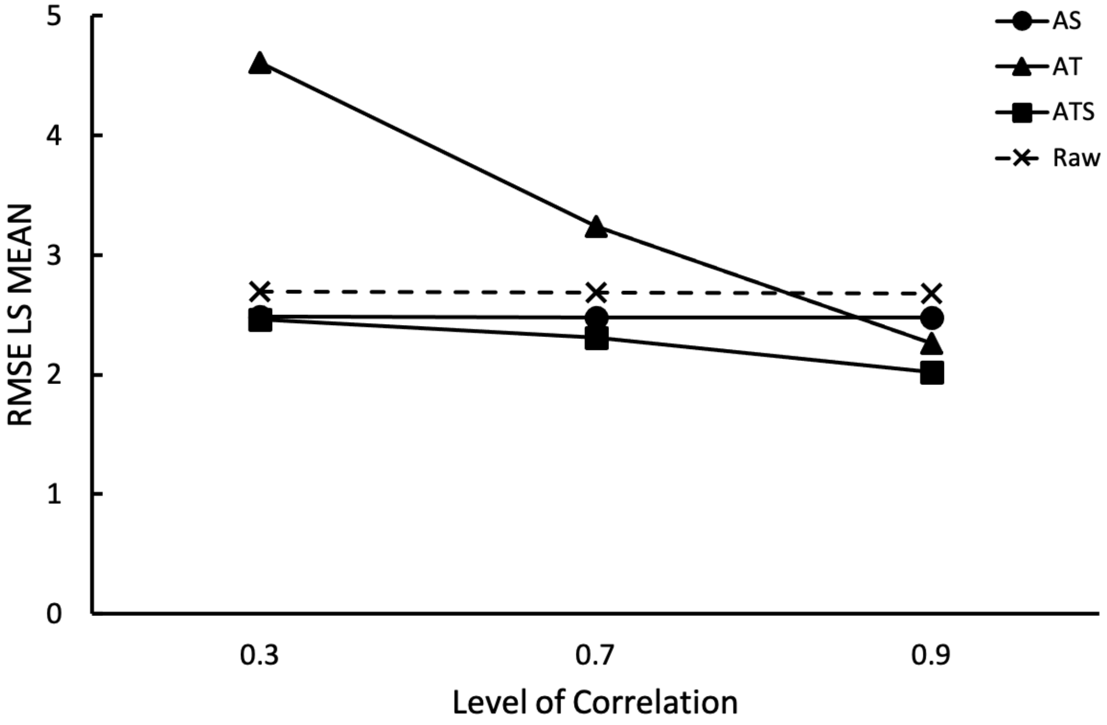

Figure 2. Average RMSE LS Mean for Subscore Estimates when K=8

For both four- and eight-item subscales, there is an interaction between the inter-subscore correlation and inclusion of the total score. That is, for AT and ATS scores, the RMSE LS Mean depends on the inter-subscore correlation. Additionally, for both four- and eight-item cases, the LS Mean is constant across inter-subscore correlations for AS scores. This is expected, given that the only weight in the AS score is coefficient α, which is unaffected by the inter-subscore correlation. These results indicate that AT scores perform poorly with respect to estimating the true subscore when subscores are correlated 0.3. Additionally, when there are eight items per subscale, the AT scores perform poorly when the inter-subscore correlation is 0.7.

RMSE of Reliability Estimates

The number of dimensions and level of correlation between subscores does not affect subscore reliability estimates and therefore those factors were averaged over within each number of items (K) condition. The average estimated reliability is 0.74 for four-item subscales and 0.85 for eight-item subscales. These values were very close to the generating values, with RMSEs of 0.006 and 0.002 for four-item subscales and eight-item subscales, respectively.

The reliability of the total score was not a directly manipulated variable in this simulation. What we consider generating values for the total score reliability are the squared correlations between the sum of the observed subscores and the sum of true subscores. Descriptive statistics for these generating values are presented by relevant condition in Table 5.

Table 5

Descriptive Statistics for the Generating Total Score Reliability Values

K	D	R	Mean (SD)	Median	Minimum	Maximum
4	2	0.3	0.79 (.03)	0.79	0.68	0.86
4	2	0.7	0.84 (.02)	0.84	0.77	0.89
4	2	0.9	0.85 (.02)	0.85	0.77	0.89
4	5	0.3	0.86 (.02)	0.87	0.79	0.89
4	5	0.7	0.92 (.01)	0.92	0.88	0.95
4	5	0.9	0.93 (.01)	0.93	0.90	0.95
8	2	0.3	0.88 (.02)	0.88	0.84	0.92
8	2	0.7	0.91 (.01)	0.91	0.88	0.93
8	2	0.9	0.92 (.01)	0.92	0.89	0.94
8	5	0.3	0.93 (.01)	0.93	0.90	0.95
8	5	0.7	0.96 (.01)	0.96	0.94	0.97
8	5	0.9	0.96 (.00)	0.96	0.95	0.97

K = Items per Subscale, D = Dimension, R = Level of Inter-subscore Correlation

Figure 3 depicts the average RMSE of total score reliability estimates for each of the 12 relevant design conditions. The best recovery occurs when there are five subscales with eight items each, and the level of correlation between them is 0.9. The worst recovery occurs when there are two subscales with four items each and the level of correlation between them is 0.3.

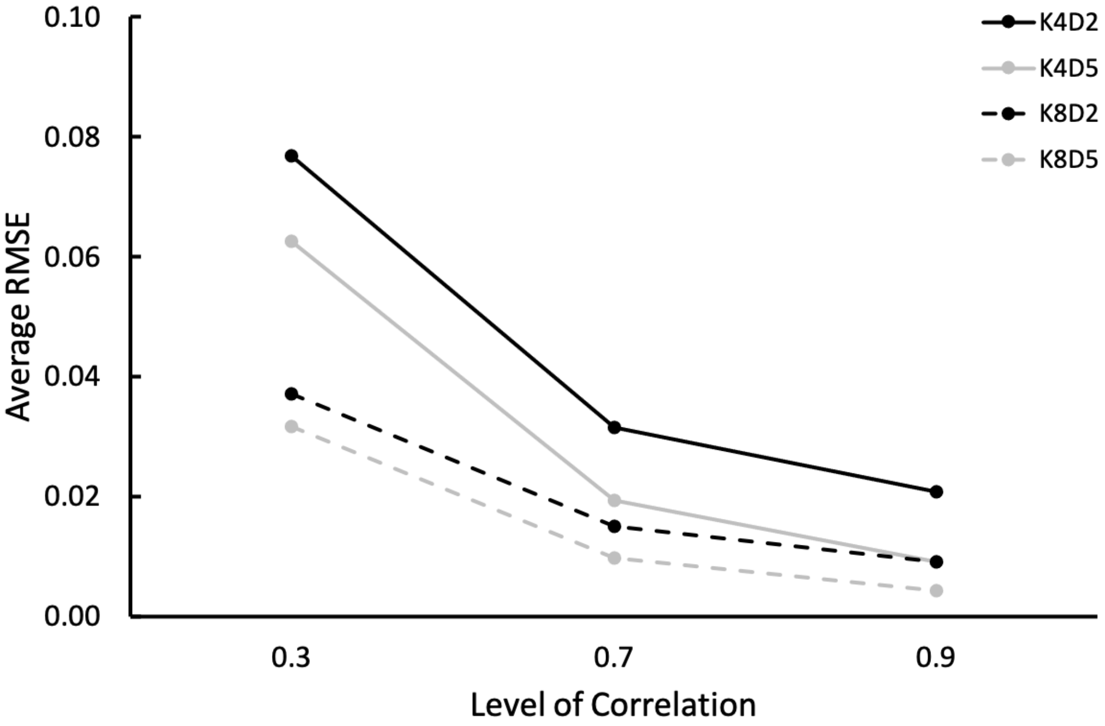

Figure 3. Average RMSE of Total Score Reliability Estimates

Bias of Reliability Estimates

The average bias for subscore reliability estimates when there are four items per subscales is -0.005; when there are eight items per subscale the average bias is -0.002. Across all conditions, subscore reliability estimates underestimated the true subscore reliability. The largest amount of downward bias occurs when there are two, four-item subscales that are correlated 0.3. The least amount of bias occurs when there are five, eight-item subscales that are correlated 0.9. Table 6 show the average bias for the total score reliability estimates. For convenience, the results for total score are also plotted in Figure 4.

Table 6

Average Bias for the Total Score Reliability Estimates per Condition

K	D	R	Mean Bias
4	2	0.3	-0.07
4	2	0.7	-0.03
4	2	0.9	-0.01
4	5	0.3	-0.06
4	5	0.7	-0.02
4	5	0.9	-0.01
8	2	0.3	-0.04
8	2	0.7	-0.01
8	2	0.9	0.00
8	5	0.3	-0.03
8	5	0.7	-0.01
8	5	0.9	0.00

K = Items per Subscale, D = Dimension, R = Level of Inter-subscore Correlation

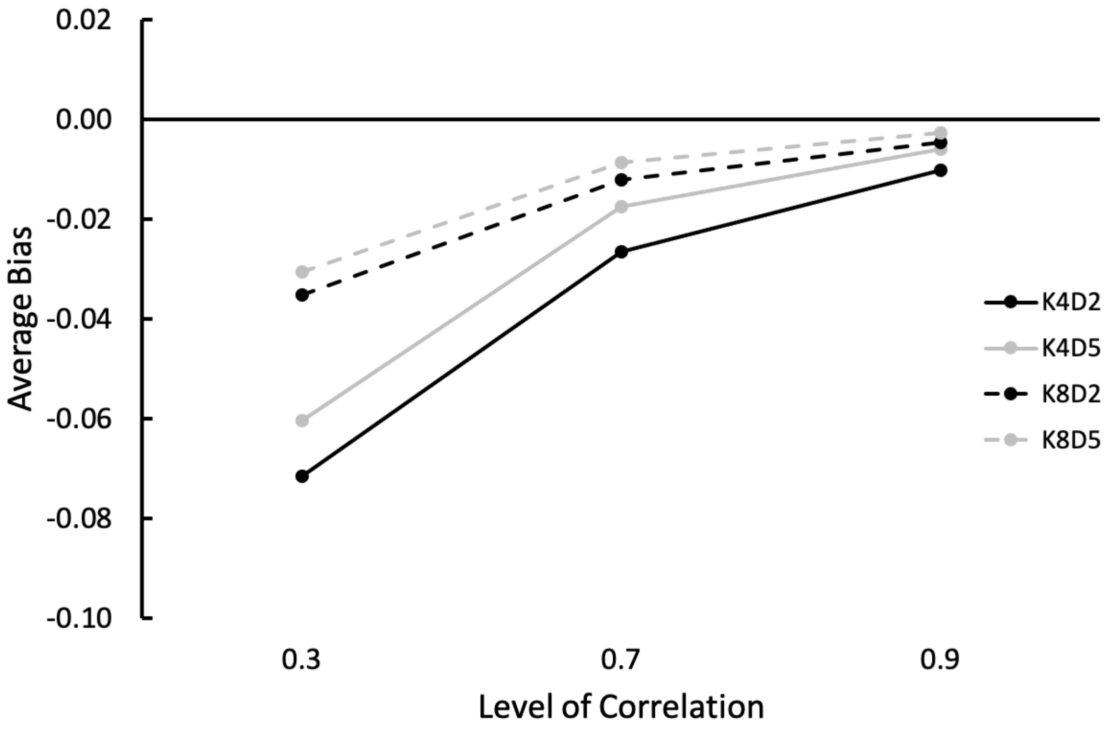

Figure 4. Average Bias of Total Score Reliability Estimates

Across all conditions, again we see that the total score reliability estimates are negatively biased. The conditions in which the level of inter-subscore correlation is 0.3 display the greatest degree of underestimation, and an inter-subscore correlation level of 0.9 displays the least amount of bias.

RMSE of Inter-subscore Correlation Estimates

Tables 7 and 8 display descriptive results for the RMSE of raw and disattenuated inter-subscore correlation estimates, respectively. The best recovery of raw inter-subscore correlations is when there are two subscales with eight items each and the level of correlation between them is 0.3. Conversely, the worst recovery occurs when there are two subscales with four items each and the subscores are correlated 0.9. In general, the higher the true correlation between subscores, the larger the RMSE. The best recovery of disattenuated inter-subscore correlations is when there are two subscales with eight items each and the level of inter-subscore correlation is 0.9. Conversely, the worst recovery occurs when there are five subscales with four items each and the subscores are correlated 0.3. In general, the higher the true correlation among subscores, the larger the RMSE. In all but one case (eight items, five subscales, 0.3 inter-subscore correlation), the disattenuated estimates have lower RMSE than the raw estimates. In the cases of the 0.7 and 0.9 inter-subscore correlations, the disattenuated correlations have RMSEs that are 64% to 85% smaller than their raw counterparts.

Table 7
Average RMSE for Raw Inter-subscore Correlation Estimates

K	D	Level of Correlation		
		0.3	0.7	0.9
4	2	0.09	0.18	0.24
4	5	0.09	0.18	0.23
8	2	0.05	0.1	0.13
8	5	0.05	0.11	0.13

K = Items per Subscale, D = Dimension

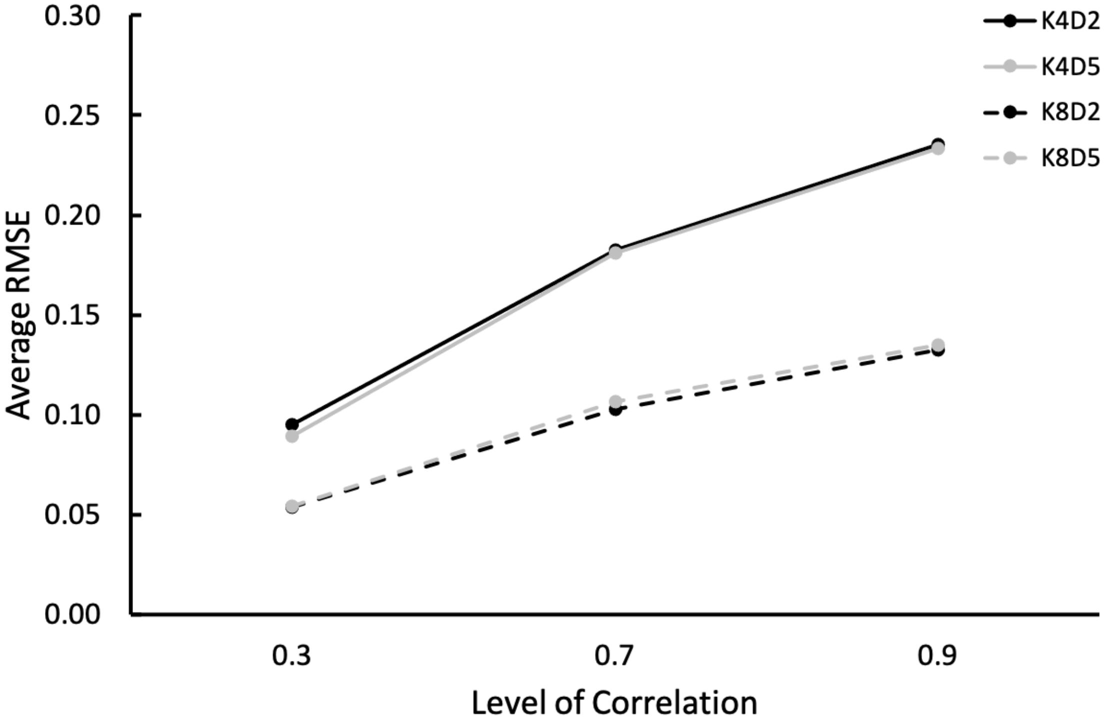

Figure 5. Average RMSE of Raw Inter-subscore Correlation Estimates

Table 8
Average RMSE of Disattenuated Inter-subscore Correlation Estimates

K	D	Level of Correlation		
		0.3	0.7	0.9
4	2	0.06	0.05	0.04
4	5	0.08	0.06	0.04
8	2	0.03	0.03	0.02
8	5	0.07	0.04	0.03

K = Items per Subscale, D = Dimension

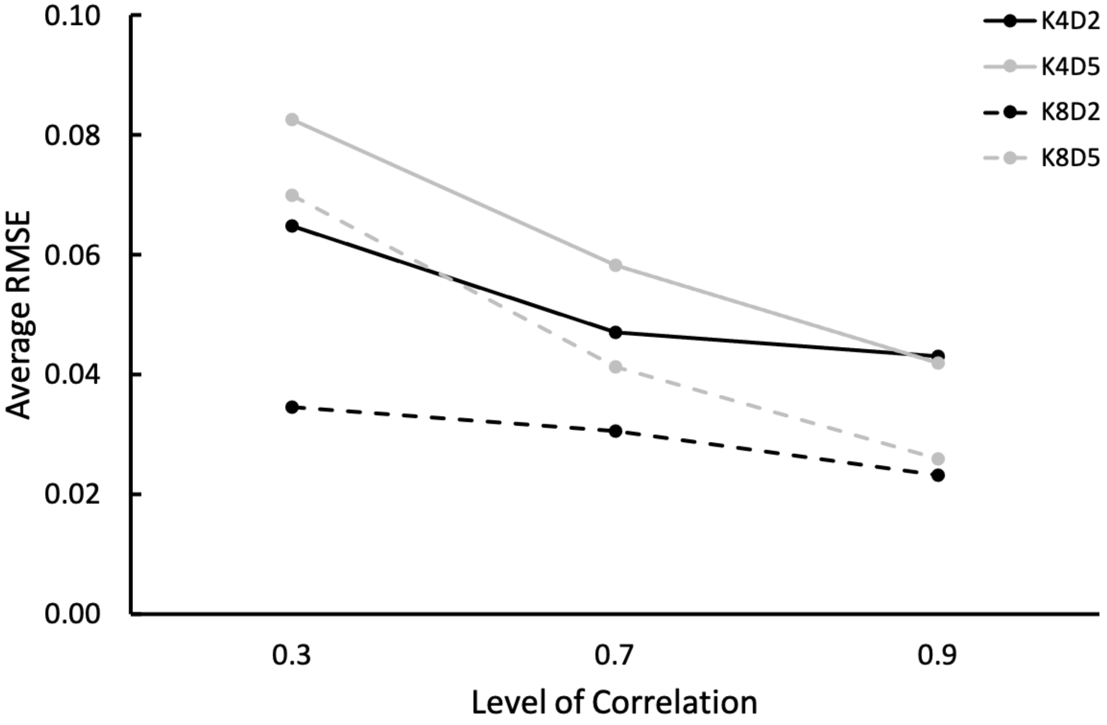

Figure 6. Average RMSE of Disattenuated Inter-subscore Correlation Estimates

Bias of Inter-subscore Correlation Estimates

Table 9 and Figure 7 display the average bias of raw inter-subscore correlation estimates, collapsing over the number of dimensions. Table 10 displays the average bias of disattenuated inter-subscore correlation estimates per condition.

Table 9

Average Bias of Raw Inter-subscore Correlation Estimates

K	R	Average Bias
4	0.3	-0.08
4	0.7	-0.18
4	0.9	-0.23
8	0.3	-0.04
8	0.7	-0.10
8	0.9	-0.13

K = Items per Subscale, R = Level of Correlation

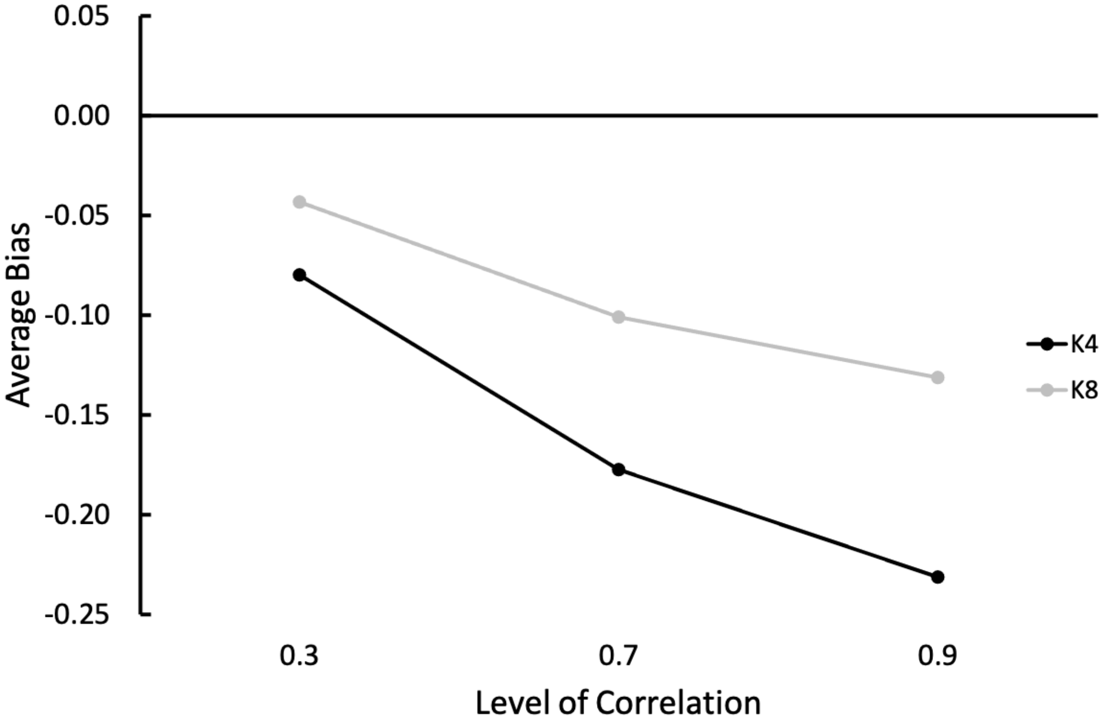

Figure 7. Average Bias of Raw Inter-subscore Correlation Estimates

Across all condition, the raw inter-subscore correlations are underestimated. The greatest degree of underestimation occurs when the true correlation among subscores is 0.9 and there are four items per subscale.

Table 10

Average Bias of Disattenuated Inter-subscore Correlation Estimates

K	D	R	Average Bias
4	2	0.3	-0.008
4	2	0.7	0.000
4	2	0.9	0.006
4	5	0.3	0.002
4	5	0.7	0.007
4	5	0.9	0.004
8	2	0.3	0.001
8	2	0.7	0.005
8	2	0.9	0.003
8	5	0.3	0.002
8	5	0.7	0.002
8	5	0.9	0.001

K = Items per Subscale, D = Dimension, R = Level of Correlation

In general, the disattenuated inter-subscore correlations are slightly overestimated. The exception is when there are two, four-item subscales and the correlation among them is 0.3, in this case, the disattenuated inter-subscore correlation is underestimated. However, when there are five, four-item subscales whose correlation is 0.3, the disattenuated inter-subscore correlation is overestimated.

MSECs

Average MSECs are presented by relevant condition in Table 11. With respect to estimating the true subscore, the observed subscore has the greatest reduction in MSE compared to the total score when the level of correlation between subscales is 0.3. Compared to both the observed subscore and the total score, the ATS score shows the greatest reduction in MSE. When there are eight-item subscales and the level of correlation between them is 0.3, the AT score performs worse than the raw (observed) subscore.

Table 11

Average Change (Reduction) in MSE Relative to the Denominator

K	D	R	$\text{MSEC}_{S/T}$	$\text{MSEC}_{AS/S}$	$\text{MSEC}_{AT/S}$	$\text{MSEC}_{ATS/S}$	$\text{MSEC}_{AS/T}$	$\text{MSEC}_{AT/T}$	$\text{MSEC}_{ATS/T}$
4	2	0.3	0.61	0.59	0.46	0.73	0.84	0.79	0.90
4	2	0.7	0.56	0.58	0.54	0.66	0.82	0.80	0.85
4	2	0.9	0.52	0.60	0.66	0.69	0.81	0.84	0.85
4	5	0.3	0.86	0.59	0.49	0.80	0.94	0.93	0.97
4	5	0.7	0.83	0.59	0.54	0.72	0.93	0.92	0.95
4	5	0.9	0.81	0.59	0.73	0.77	0.92	0.95	0.96
8	2	0.3	0.60	0.39	-0.18	0.60	0.76	0.54	0.84
8	2	0.7	0.55	0.38	0.13	0.49	0.72	0.61	0.77
8	2	0.9	0.52	0.38	0.44	0.53	0.70	0.73	0.77
8	5	0.3	0.86	0.38	-0.20	0.71	0.91	0.83	0.96
8	5	0.7	0.83	0.39	0.02	0.57	0.90	0.84	0.93
8	5	0.9	0.81	0.38	0.50	0.64	0.88	0.91	0.93

K = Items per Subscale, D = Dimension, R = Level of Inter-subscore Correlation

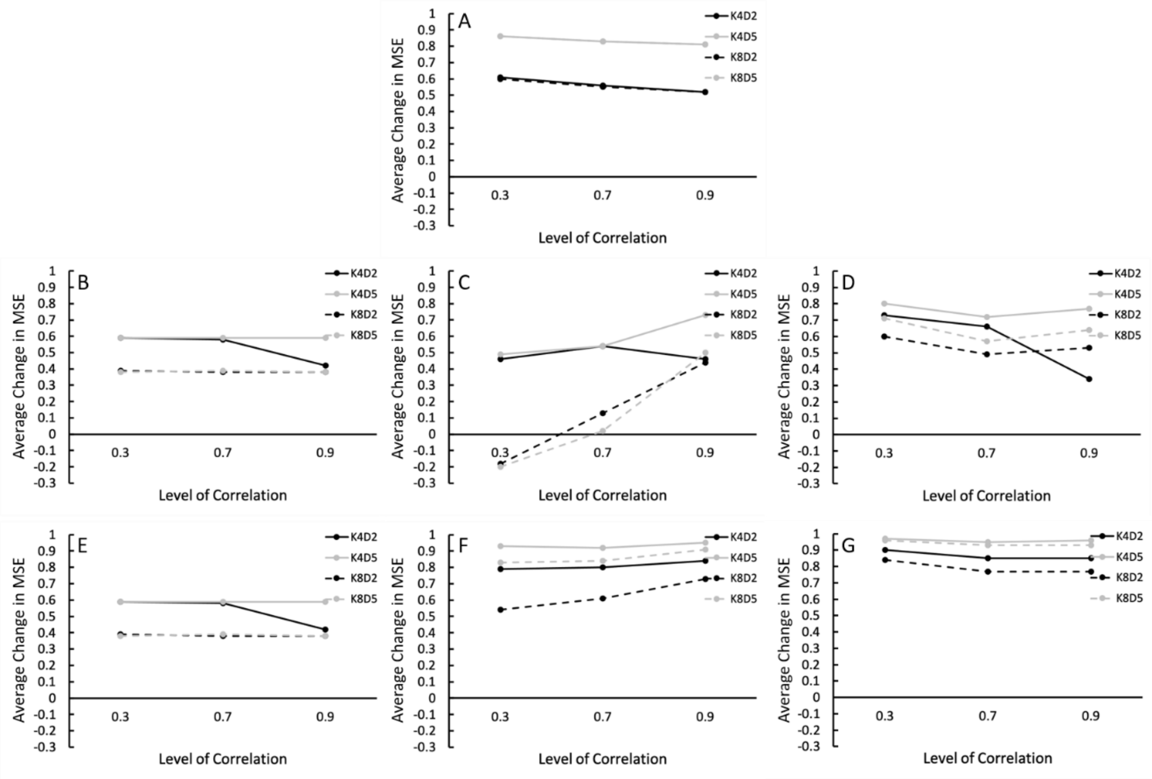

Figure 8. Average Change (Reduction) in MSE Relative to the Denominator. $MSEC_{S/T}$ *(A);* $MSEC_{AS/S}$ *(B);* $MSEC_{AT/S}$ *(C);* $MSEC_{ATS/S}$ *(D);* $MSEC_{AS/T}$ *(E);* $MSEC_{AT/T}$ *(F);* $MSEC_{ATS/T}$ *(G)*

CHAPTER 4

DISCUSSION

In this section we first discuss results from the three specific hypotheses that we examined. Next, we compare our findings to Sinharay (2010). Then we address the quality of the reliability and inter-subscore correlation estimates, and revisit validity and dimensionality. We conclude by offering several take-away messages regarding the use of subscores and subscore augmentation techniques in the psychological and health sciences.

Revisiting the hypotheses

It was hypothesized that observed subscores would always be a more accurate estimate, as judged by RMSE, of the true subscore than the total score, and it was indeed the case—at least for these simulation conditions—that the observed subscores were always more accurate than the total score. This suggests that even when the latent correlation among subscores is 0.9, the observed subscore is able to capture more of the true subscore than the total score. Thus, even if the estimated disattenuated inter-subscore correlation is ~0.9, researchers ought to use subscores instead of total scores. However, comparing across the levels of inter-subscore correlation, the smallest gain in prediction (over the total score) occurs when the subscores are correlated 0.9.

Additionally, it was hypothesized that the adjusted subscores would always be more accurate than the observed subscore. In the cases of the AS and ATS subscores, this was in fact true. This was not the case for the AT subscores, as evidenced by the negative $MSEC_{AT/S}$ values in Table 11. In conditions with eight items per subscale when the subscales are correlated 0.3, the raw subscore is more accurate than the AT score. Extensive efforts were undertaken to verify that the simulation and analysis code were functioning as intended and no errors were discovered. For context, it is useful to note that this simulation considered lower inter-subscore correlations (r=0.3) than had previously been studied. For example, the lowest value included in Sinharay (2010) was 0.7. It was at this lowest level of correlation among subscores where the AT scores failed to outperform their non-augmented counterparts. To better understand why this result occurred, we looked at the weight on the total score in the ATS score calculations in comparison to the weight in the AT score calculation. These aren't directly comparable, as the ATS weight is conditional on the subscore component, but examined how much the weight on the total score dropped in the AT and ATS cases as the inter-subscore correlations decreased. For example, the average weight on the total score for the ATS score calculation in the conditions with eight items and five dimensions were 0.12, 0.07, and 0.03 for inter-subscore correlations of 0.9, 0.7, and 0.3, respectively. The average weight on the total score for the AT score calculation in the same cells were 0.193, 0.191, and 0.185. It is suggestive that the weights decrease more quickly as the correlation among subscores decreases in the ATS case than in the AT case. It appears that, for reasons not yet understood, the AT weight does not adjust downward enough to account for a total score of limited predictive utility.

In terms of change in MSE, it was hypothesized that the ATS scores would always be the most accurate estimate of the true subscore. This was found to be true in all conditions. This is unsurprising given that the ATS score contains the most information. The largest change in MSE from using the ATS score compared to the raw score, ($MSEC_{ATS/S}$), occurs when there are five subscales, correlated 0.3, with four items each. The average estimated reliability was 0.74 for four-item subscales and 0.85 for eight-item subscales. The pattern of results displayed in Table 11 show us that observed scores with reliability averaging 0.74 benefit the most from augmentation. Scores with reliability averaging 0.85 also benefit from augmentation, but not as much as the less reliable scores. This makes sense as less reliable scores have more room for improvement.

Looking at Table 11, some informative patterns emerge. First, $MSEC_{AS}/_{S}$ doesn't change across levels of correlation among subscores like $MSEC_{AT}/_{S}$ and $MSEC_{AS}/_{S}$ because the AS score is unaffected by inter-subscore correlation. The weight in the AS score is coefficient α, which is affected by items per subscale, and this pattern is made evident in the table when moving from four to eight items per subscale. Within the four- and eight-item cases, the MSEC is uniform. However, comparing all of the four-item cases to all of the eight-item cases, there is greater improvement with four-item subscales. This is because there is more room for improvement when you start with scores that aren't very reliable (i.e., shorter). There is more to be gained in terms of accuracy when you're starting with shorter subscales and moving from a raw subscore to the AS score.

Another interesting pattern is that within $MSEC_{AT}/_{S}$ and within $MSEC_{ATS}/_{S}$ the same repeating pattern emerges as you move down the rows (each row represents a simulation condition). Even though $MSEC_{AT}/_{S}$ changes as a function of items per subscale, number of subscales, and level of inter-subscore correlation, the most noticeable pattern is that as the level of inter-subscore correlation increases, there is more improvement in using the AT score versus the raw subscore. This make sense because the subscores are conveying less and less unique information as they become more correlated with each other, and the AT score is weighted by a regression coefficient that accounts for the correlation between the true subscore and observed total score. The higher that correlation, the more weight that is placed on the total score in the AT calculation. Subsequently, the AT score shows more improvement (more change) over the raw subscore as the inter-subscore correlation increases. Additionally, there appears to be an interaction between the number of items per subscale and level of correlation among subscores such that the $MSEC_{AT/_{S}}$ *depends* on the level of those two things. When there are four items per subscale, there is always benefit in using the AT score compared to the raw subscore, regardless of level of correlation among subscores. But when there are eight items per subscale and the level of correlation among subscores is 0.3, the raw subscore is actually better than the AT score at predicting the true subscore. This finding has important implications. Given that subscale length and inter-subscore correlation are proxies for reliability and distinctiveness, respectively, then it can be concluded that raw subscores are more accurate than the AT score when the raw scores are distinct (r ~0.3) and reliable (α ~ 0.85). Both conditions must be present. This finding corroborates with Haberman and Sinharay in that subscores must be distinct and reliable to have added value over the total score. However, they compared what is effectively the AT to the AS score, relative to the subscore sample mean, whereas we compared the AT score to the raw subscore directly.

$MSEC_{ATS/_{S}}$ shows an interesting and informative pattern of results. Looking at how $MSEC_{ATS/_{S}}$ changes as a function of inter-subscore correlation, we see that in every case the MSEC starts high when the inter-subscore correlation is 0.3, drops when inter-subscore correlation is 0.7, but then increases when inter-subscore correlation is 0.9 (although not as high a value as when inter-subscore correlation is 0.3). This likely reflects a tradeoff between distinctiveness and reliability. The lower the correlation between subscores, the more unique information there is in the total score as a second predictor. According to these results, though, it is also the case that the total score is less reliable when the inter-subscore correlations are low. As the inter-subscore correlations increase, so does the total score reliability, but because the subscores are more correlated there is less unique information the total score can bring to ATS prediction. However, the $MSEC_{ATS}/_{S}$ when the subscores are correlated 0.9 is never greater than the $MSEC_{ATS}/_{S}$ when they are correlated 0.3. This suggests two things: (1) according to these simulation conditions, distinctiveness has a greater impact on the value of the ATS score than does reliability and (2) the tradeoff between total score distinctiveness and reliability is slightly more complex than we might have been anticipated.

Lastly, from Table 11 we can see that $MSEC_{ATS}/_{S}$ values are uniformly larger/higher than $MSEC_{AT}/_{S}$ values. Given the simulation design, we expected the ATS scores would be the most accurate. From a regression perspective, the two predictor variables in the ATS score are the single predictor variables from the AS and AT scores. If there is any unique covariance between the predictors and the outcome (in this case, the ATS score), then the joint prediction must account for more variance and subsequently have a lower MSE. A potentially more informative way to look at these results is to examine which design features lead AS or AT to be closer to ATS.

How do these results compare to Sinharay (2010)?

The most substantial difference between the current study's findings and Sinharay's (2010) findings is that, compared to education, it will be more common to have subscores that have added value in psychology and health applications. Sinharay (2010) posits that the added value of subscores depends on the interaction between the subscore reliability and the level of correlation among subscores. The conclusion that subscores need to be sufficiently distinct and reliable to have added value generalizes across the two studies, but it is easier to obtain sufficiently distinct and reliable subscores in scales that mimic what we commonly see in psychology and health. This result is important because it follows up on a recommendation from Sinharay (2010) that further research ought to consider polytomous items. The conditions of this study match Sinharay's call for additional research and the results suggest that conditions commonly found in behavioral and health assessment will lead to useful subscores on a regular basis.

Another difference between the current study's results and Sinharay's (2010) results was the number of items required to achieve adequate reliability. In Sinharay's (2010) study, subscales with 20 items produced scores with reliability estimates ranging between 0.72 to 0.77. He recommends having a minimum of 20 items per subscale to have any hope of having subscores with added value. In contrast, results from the current study indicate that reliability of ~0.74 can be obtained with as few as four items per subscale. Given that polytomous items typically have higher slopes than dichotomous items, it is not surprising that fewer items were needed in the current study to achieve approximately the same level of reliability.

For subscales to be considered *sufficiently distinct* from one another, Sinharay's (2010) results suggest that the disattenuated correlation among subscores needs to be less than 0.85. Our results suggest that, for the conditions in this simulation, even when the true correlation among subscores is as high as 0.9, there is still benefit in using the raw subscore over the total score as an estimate of the true subscore. In practice though, people often ignore subscores altogether and therefore using the total score <u>as an estimate of the true subscore</u> is not common. Although these results are not directly comparable to Sinharay's due to the use of different methods and dependent variables in our simulations, it is still interesting to note that in psychology and health contexts, the subscores can be highly correlated and still measuring distinct constructs.

Quality of Reliability and Inter-subscore Correlation Estimates

It was important to assess the recovery of the reliability and inter-subscore correlation estimates as poor recovery of these quantities could have led to poor performance of the augmentation procedures. In terms of RMSE, both subscore and total score reliability estimates showed excellent recovery. This is consistent with existing literature that suggests good recovery with sample sizes of 300 (Nunally & Bernstein, 1984). We used coefficient α as an estimate of reliability in this study, but there are arguments against using α with multidimensional scales (see Schmitt, 1996). Kamata, Turhan, and Darandari (2003) argued that stratified α may be a more accurate estimate of reliability on multidimensional scales. However, Sinharay (2010) computed stratified α for some of his operational data and found that the values were very close to coefficient α, leading him to report results using coefficient α. Given these findings, and the fact that α is still a very common estimate of reliability, we felt using α in this study was a defensible choice.

Due to unreliability in the scores, the raw inter-subscore correlations were underestimated relative to the generating correlations between the dimensions across all conditions. The RMSE and average bias for the raw inter-subscore correlations were much larger for inter-subscore correlations of 0.7 and 0.9 than for 0.3. For example, four-item subscales that had a true correlation of 0.9 were underestimated by 0.23 on average. This amount of bias is expected according to Fan (2003). The disattenuated inter-subscore correlations performed much better, indicating that the correction for attenuation due to unreliability worked as expected. Sinharay (2010) pointed out that the average disattenuated correlation among subscores was always very close to the true level of correlation among subscores. In the current study, we also had excellent recovery of the disattenuated correlations among subscores.

Revisiting Validity and Dimensionality

From a statistical perspective, we found that the ATS scores are the most accurate estimate of the true subscore in terms of MSE. In addition, given modern software capabilities, they are easy to compute. Therefore, we broadly suggest the use of the ATS score. However, the AS score should not be overlooked. It also shows improvement over the observed subscore and it has the benefit of being easier to compute. This is not to suggest that people settle for less accurate predictions by using the AS score instead of the ATS score. But from a practical perspective, the AS score may be easier for psychologists to implement and it will still be an improvement over the raw subscore. Furthermore, from a validity perspective one might consider the AS score a "more pure" estimate of the true subscore over the ATS score because the subscore is the only predictor variable in the AS score whereas the ATS score contains both the subscore and total score for predictor variables, potentially rendering the interpretation of the ATS score more difficult.

However, there is one situation in which augmented scores should not be used and that is when the scores are being used for competitive purposes. Wainer et al (2001) and Edwards (2006) use the Olympic 100-m sprint to illustrate this point. The gold medal is awarded to the person who runs the fastest *that day*, regardless of how fast they ran on previous days and regardless of how fast they will run in future races. All that matters in the Olympics (and most competitive sports) is how well you perform *that day*. Therefore, it is the general consensus among Wainer et al (2001), Edwards (2006), and the current study that when scores are to be used for competitive purposes, augmented scores ought not to be used. However, even this delineation is complex because the competition could be based on the day-of performance (such as in all competitive sports) or it could be based on what the scores indicate/predict about future performance (such as competing for scholarships). Wainer et al (2001) argued that if it is the latter, then augmented scores (i.e., estimates that utilize all information available) should be used because they are the most accurate. In general, if scores are being used for measurement purposes rather than contest purposes then we suggest the use of augmented scores, either the AS or the ATS score. The potential concerns that arise with using these scores are discussed next.

The procedure to calculate the AS score involves regressing observed scores towards the sample mean proportional to the unreliability of the scores. The potential ethical concern that arises in this procedure pertains to which sample mean the score is being regressed to. If everyone in the sample is assumed to be exchangeable then using one overall sample mean will avoid this problem. However, there may be substantive differences within the sample that the researcher/clinician wants to account for. The researcher could accomplish this by calculating separate means for each subgroup in the sample and then use those means to calculate the adjusted scores. One area where this may be applicable is in health and medical research, where one's race, gender, or ethnicity may be correlated with the outcome measure (e.g., diagnosis, prognosis, disease propensity). For example, women have a higher rate of breast cancer diagnoses than men, so it seems apt to factor in gender as a grouping variable when trying to determine a patient's propensity for a breast cancer diagnosis. In this case, considering group membership information is appropriate and does not raise ethical concerns (to our knowledge).

However, there are also situations in which there may be substantive differences among the sample but to account for those differences with separate means would be considered inappropriate. For example, female examinees tend to score higher on the Verbal portion of the SAT compared to male examinees, such that a more accurate estimate of the examinee's proficiency could be obtained if their gender were factored in when calculating their adjusted score. Although a more accurate estimate of an examinee's proficiency could be obtained if we considered the examinee's gender, it would be unfair and unethical to change an individual's score because of their gender, especially if the score is being used in a contest (e.g., to award a scholarship).

As mentioned previously, augmented subscores have been criticized for lacking validity (e.g., Skorupski & Carvajal, 2010; Stone et al., 2010). Specifically, there has been concern that once a subscore has been augmented, it no longer measures the construct of interest in a valid way. The ATS score contains the total score predictor variable, which contains information that is outside of the subscore. By pulling in external information from the total score we are changing the subscore, and in this way there is room for potential threats to validity. Part of this concern stems from the fact that the correlations among the adjusted subscores can be quite high, therefore obscuring the interpretation of the subscore (Skorupski & Carvajal, 2010; Stone et al., 2010). However, Sinharay et al (2011b) pointed out that the correlations among the adjusted subscores will *always* be higher than the unadjusted subscores because the adjusted subscores share a common component: the total score. Therefore, the concern that the meaning of the subscore is obscured due to the higher correlations among the adjusted subscores is partially taken care of by recognizing that the higher correlations among subscores are due to a mathematical fact. An additional concern is that part of what defines a construct is how it relates to other constructs. By changing how the constructs relate to one another (i.e., convergent and discriminant validity), we've changed the construct and introduced potential validity concerns.

General Advice for Applied Users

Broadly speaking, if a scale is designed to yield subscores, use them. Outside of competitive contexts, to increase the accuracy of true score estimation and prediction, use adjusted scores. For conditions like the ones we simulated, we suggest the use of the ATS score or Wainer et al's (2001) augmented subscore because they are the most accurate in terms of MSE. However, if users are uncomfortable with using the ATS score for validity reasons, then we suggest the use of the AS score (i.e., Kelley's regressed estimate, posterior mean in Bayesian analysis). These scores will always be an improvement over the observed subscore and are easy to compute. The AS score is only correcting for unreliability, while the ATS score is correcting for unreliability *and* pulling in external information from the total score. We anticipate some pushback to these suggestions. For one, users may argue that there is no meaningful payoff. While the adjusted scores are clearly preferable when considering RMSE, more work is needed to find practical effect sizes that translate these results for applied researchers. The second criticism of adjusted scores is that they have a weaker validity case. Sinharay et al (2011b) has shown that this is not the case, and even demonstrated the situation in which adjusted subscores *do* lack validity: adjusted subscores lack validity when they are highly correlated with the total score and have low reliability.

Limitations and Future Directions

As with any simulation study, this one only generalizes to the conditions of the study. Our simulation did not incorporate any model error, which is not a condition we would expect to find in the real world. Another limitation is that we did not directly control/manipulate reliability, but rather focused on two scale lengths that we felt were realistic for the fields we were interested in focusing on. This led to a perfect confound between scale length and reliability. In future studies we would like to vary the generating slope distributions so that we can look at the impact of reliability separate from scale length on the added value of adjusted scores. As part of this study, we intended to evaluate the performance of the VAR simplification as well as consider the utility of VAR in communicating with applied researchers. After considering this question, we determined that VAR offered no conceptual advantages over the MSECs considered here. Despite this, we are not satisfied that MSEC is the best metric to convey the impact of augmentation procedures to potential users.

One surprising finding of the current study was that when the correlations among the subscores were low (e.g., 0.3), the AT scores were often *worse* than the raw subscores. We had expected all the augmented subscores to outperform their raw counterparts for all simulated conditions. There is a large gap in studied inter-subscore correlations (0.3 to 0.7) and it would be worthwhile to fill in this gap to determine at what correlation among subscores the AT scores cease outperforming the raw subscores. It could also be that the current estimates of the weights being used are inaccurate or suboptimal in these low inter-subscore correlation cases. More work is needed to understand if the computation of the AT scores can be improved, or if we simply must avoid considering them at some inter-subscore correlation-based cutoff.

Conclusions

The current study had two overall objectives: to assess when subscores should not be ignored in favor of the total score, and to examine how much value there is in using an augmented subscore over a non-augmented subscore in scales that mimic what we commonly see in psychological and health contexts. In general, when a scale is designed to yield subscales, we recommend using them (even if they are moderately to strongly correlated with one another). In terms of MSE and according to the results of the current study, we found that the AS and ATS scores reflect a more accurate estimate of the true subscore than the observed subscore.

REFERENCES

American Educational Research Association, A. P. A., & National Council on Measurement in Education. (2014). Standards for validity. In *Standards for Educational and Psychological Testing* (pp. 23-32). American Educational Research Association.

Birnbaum, A. (2008). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & Novick, M. R. (Ed.), *Statistical Theories of Mental Test Scores* (pp. 395-480). Information Age Publishing Inc.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289.

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

Edwards, M. C., & Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow? *Journal of Educational and Behavioral Statistics, 31*(3), 241-259.

Fan, X. (2003). Two approaches for correcting correlation attenuation caused by measurement error: Implications for research practice. *Educational and Psychological Measurement, 63*(6), 915-930.

Feinberg, R. A. (2012). *A simulation study of the situations in which reporting subscores can add value to licensure examinations* [Doctoral dissertation, University of Delaware]. ProQuest Digital Dissertations database.

Feinberg, R. A., & Jurich, D. P. (2017). Guidelines for interpreting and reporting subscores. *Educational measurement: Issues and practice, 36*(1), 5-13.

Feinberg, R. A., & Wainer, H. (2014). A simple equation to predict a subscore's value. *Educational measurement: Issues and practice, 33*(3), 55-56.

Finch, H., & Habing, B. (2007). Performance of DIMTEST- and NOHARM-Based Statistics for Testing Unidimensionality. *Applied Psychological Measurement, 31*(4), 292-307. https://doi.org/10.1177/0146621606294490

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

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*(2), 204-229. https://doi.org/10.3102/1076998607302636

Hill, C. D. (2004). *Precision of parameter estimates for the graded item response model* [Unpublished master's thesis]. The University of North Carolina at Chapel Hill.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6*(3), 153-160.

Kamata, A., Turhan, A., & Darandari, E. (2003). Estimating reliability for multidimensional composite scale scores. annual meeting of American Educational Research Association, Chicago, IL,

Kelley, T. L. (1923). *Statistical method*. Macmillan.

Lord, F. M., & Novick, M. R. (2008). *Statistical Theories of Mental Test Scores*. Information Age Publishing Inc.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological methods, 23*(3), 412.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling.

Reckase, M.D. (2007). Multidimensional item response theory. In Rao, C.R., & Sinharay, S. (Eds.), *Handbook of statistic.* (Vol. 26, pp. 607–642). Amsterdam: North-Holland.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological assessment, 8*(4), 350.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47*(2), 150-174.

Sinharay, S., Haberman, S., & Boughton, K. (2015). Too simple to be useful: A comment on Feinberg and Wainer (2014). *Educational measurement: Issues and practice, 34*(3), 6-8.

Sinharay, S., Haberman, S. J., & Wainer, H. (2011). Do adjusted subscores lack validity? Don't blame the messenger. *Educational and Psychological Measurement, 71*(5), 789-797. https://doi.org/10.1177/0013164410391782

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational measurement: Issues and practice, 30*(3), 29-40.

Skorupski, W. P., & Carvajal, J. (2010). A comparison of approaches for improving the reliability of objective level scores. *Educational and Psychological Measurement, 70*(3), 357-375.

Stanford, M. S., Mathias, C. W., Dougherty, D. M., Lake, S. L., Anderson, N. E., & Patton, J. H. (2009). Fifty years of the Barratt Impulsiveness Scale: An update and review. *Personality and individual differences, 47*(5), 385-395.

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2009). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education, 23*(1), 63-86.

Wainer, H., & Thissen, D. (2009). True score theory: The traditional method. In D. Thissen, & Wainer, H. (Ed.), *Test Scoring* (pp. 23-72). Routledge.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., Swygert, K. A., & Thissen, D. (2009). Augmented scores—"Borrowing strength" to compute scores based on small numbers of items. In D. Thissen, & Wainer, H. (Ed.), *Test Scoring* (pp. 343-387). Routledge.

Wilcox, R. R., & Tian, T. (2008). Comparing dependent correlations. *The Journal of general psychology, 135*(1), 105-112.