Subscores Aren't for Everyone: Alternative Strategies for Evaluating Subscore Utility[1]

Mark R. Raymond and Richard A. Feinberg
National Board of Medical Examiners

Background and Purpose

Conventional methods for evaluating the utility of score profiles rely on covariances (or correlations) among subscores and on traditional indices of reliability (e.g., coefficient alpha). This is true of informal methods that evaluate subscore correlations (Haladyna & Kramer, 2004), and of more formal methods proposed by Haberman and colleagues (Haberman, 2008; Sinharay, 2013) that quantify the extent to which subscores add variance that is reliably different from what can be obtained from the total score (Haberman, 2008). Feinberg and Wainer (2014) empirically derived a simplified variation of Haberman's approach and referred to it as the value added ratio (*VAR*); when *VAR* is greater than 1.0, subscores are worth reporting. Studies applying *VAR* and similar indices to data from several operational testing programs have demonstrated that subscores are seldom worth reporting (e.g., Sinharay, 2010; 2013).

Although *VAR* and other correlation-based methods for evaluating subscore utility provide a useful way to summarize relationships among variables, they have several limitations.

- The correlation is a group index and may not accurately convey the degree of subscore profile variability among low proficiency examinees or examinees who belong to certain subgroups (Sinharay & Haberman, 2014). The high correlations often observed among subscores are driven, in part, by high proficiency candidates who tend to score well in all content areas.
- Correlations ignore systematic differences in subtest (i.e., task) difficulty. Subtraction and addition are highly correlated, but are meaningfully different constructs for some people. A struggling second grader who scores 60% in subtraction and 90% in addition is poorly served by a total score of 75%. Similarly, push-ups and pull-ups are well correlated, but one is more difficult than the other. These differences in difficulty can and sometimes should be scaled away, but other times they are important to acknowledge.
- Conventional reliability coefficients and standard errors of measurement (SEM) are averages; they fail to take into consideration the well-known fact that score precision varies by examinee.
- Usual methods for evaluating subscore utility answer the question "Are these subscores reliably different from the total score?" However, examinees or other users are often more interested in the question, "Are these two subscores different from each other?" This is, admittedly, a minor limitation, because the two questions often give the same answer.
- *VAR* and related methods for evaluating subscore utility are not sensitive to the possible outcome that subscores will be informative for some individuals but not for others. They unnecessarily frame the issue as a dichotomous decision – that subscores are either useful or that they are not. It is easy to imagine situation where $VAR < 1.0$ and yet subscores are useful for some subset of examinees. Even in those few instances for which $VAR > 1.0$ and subscores are deemed useful, there undoubtedly will be individuals for whom subscores provide no meaningful information beyond the total score.

In short, *VAR* and related correlation-based indices are helpful for summarizing the relationships among variables. However, by focusing exclusively on variables instead of people, they provide an incomplete

[1] Presented at the annual conference of the National Council of Measurement in Education, April 2015, Chicago, IL.

summary of profile utility. A more effective approach to making decisions about subscores would include an analysis of actual score profiles – by focusing on rows of data rather than just the columns.

This paper introduces an alternative method for evaluating subscores. It assumes score profiles may be useful for some examinees but not for others. The method: (a) is sensitive to individual differences in profile variability; (b) accounts for differences in measurement precision across examinees; and (c) is consistent with multivariate generalizability theory. The method is applied to real test data and compared to *VAR*. The next section describes the indices and the data to which they were applied.

Method

Overview

Brennan (2001, p. 323) introduced a reliability-like index for score profiles based on generalizability theory. Designated as $\mathcal{G}$, it indicates the proportion of variance in observed score profile variance attributable to universe (true) score profile variance. $\mathcal{G}$ estimates the reliable profile variance for a population; as such it is an average, with the same value applying to all examinees in a particular sample. Of course, profile variability varies by individual, as does the amount of measurement error in the scores that make-up the scores in that profile. The following text describes an index for each individual that summarizes the amount of observed variance relative to error variance expected in a score profile.

Profile Variance and Error Variance

***Observed Score Profile Variability*.** The variability of a score profile indicates the extent to which that profile contains information about an examinee's proficiency across different content domains. Relatively flat profiles contain little or no information beyond the total score, while profiles with much dispersion carry information. While there are different ways to gauge the variability of a score profile, one traditional index is the within-examinee variance across subtests. The computation for the profile variance is made explicit here as it comes into play later:

$$\sigma^2_p = SD^2_p = \frac{\sum_s (X_{ps\bullet} - X_{p\bullet\bullet})^2}{n_s - 1} \quad , \tag{1}$$

where $X_{ps\bullet}$ is the score for examinee p on subtest s, and $X_{p\bullet\bullet}$ is the mean for examinee p across s subtest domains, and n_s is the number of subtests.

One common observation is that σ^2_p varies as a function of examinee ability, with low scoring exhibiting more variable score profiles than high scoring examinees, suggesting the possibility that score profiles might be more useful for low-scoring examinees. Another finding is that low-scores are associated with greater amounts of measurement error than high scores (e.g., Brennan, 1998; Feldt, Steffen, & Gupta, 1985). The index σ^2_p by itself is limited because measurement error contributes to profile variability, with unreliable subtests producing more variable profiles. One needs to determine if the more variable score profiles for low scoring examinees is a consequence of greater error variance.

Error Variance. Traditional subscore reliabilities and *SEM*s are constant over all examinees; they have limited use here because the *SEM* is known to vary with examinee proficiency. The conditional *SEM*, or its square, indicates the error for individual examinees and provides a useful way to determine the expected variability in score profiles due to random noise.

Brennan (1998, 2001) gives the computations for two types of conditional errors based on generalizability-theory: absolute error variance and relative error variance. Absolute error is generally greater than relative error, and is common in domain-referenced testing; it is also easier to compute. This paper is restricted to

absolute error, although methods illustrated here could apply to relative error. The absolute conditional error variance for examinee p on subtest s is given by:

$$\sigma^2(\Delta_{ps}) = \frac{\sum_i (X_{psi} - X_{ps\bullet})^2}{n_i(n_i - 1)}, \quad (2)$$

where X_{psi} is the response (0, 1) by examinee p to item i in subtest s, and $X_{ps\bullet}$ is the mean score over n_i items in subtest s for examinee p (see Brennan 2001, equation 5.32 for an equivalent expression). This equation applies to performance ratings and dichotomously-scored items, although there are simpler computations for dichotomous scores (Brennan, 2001; Lord, 1955). The square root of equation (2) is designated as the conditional SEM.

The relationship of equation (2) with the total group reliability is well known: the square root of the mean of $\sigma^2(\Delta_p)$ is the traditional total group SEM; the reliability coefficient can be obtained from the total group SEM and SD (Lord, 1955; 1957). Equation (2), which corresponds to absolute conditional error, is the basis for KR-21 in classical theory or to the phi coefficient in generalizability theory. Meanwhile, the equation for relative conditional error, designated as $\sigma^2(\delta_p)$ (Brennan, 2001), serves as the basis for KR-20, coefficient alpha, and the generalizability coefficient. We point this out because a fundamental point of this paper is that traditional total group indices can be deconstructed into their individual components.

Let $\sigma^2(\Delta_{p\bullet})$ indicate the mean error variance over subtests for examinee p:

$$\sigma^2(\Delta_{p\bullet}) = \frac{\sum_s \sigma^2(\Delta_{ps})}{n_s} \quad (3)$$

This value estimates for each examinee the expected variability across subtests just due to noise (random error). As we later illustrate, it is informative to compare mean error variances and profile variances on the same plot.

Ratio of Profile Variance to Error Variance. *PVEV* indicates the extent to which the score profile for examinee p provides information above that expected by measurement error:

$$PVEV = \sigma^2_p / \sigma^2(\Delta_{p\bullet}) \quad (4)$$

As the ratio of two variances, this index should follow an *F* distribution with an expected value of 1.0. Values substantially larger than 1.0 indicate that the score profile contains variability greater than what is expected due solely to measurement error. Although interest here is in *PVEV* for individual examinees, the mean *PVEV* obtained over numerous samples of examinees is expected to exhibit a monotonic relationship with $\mathcal{G}$, the reliability of within-person score profiles for those same samples.

Value Added Ratio

VAR represents a slight modification of the indices due to Haberman (2008). The original Haberman index is based on the notion that a subscore has value if it can predict a parallel measure of itself on a future test. If the prediction is more accurate than what can be obtained from the total test score, the subscore adds value; if the subscore is less accurate than the total score for predicting future subscore performance, then it is not useful. Haberman proposed calculating the accuracy of the subscore and the total score as predictors, and referred that index as the proportion reduction in mean-square-error (PRMSE).

Feinberg and Wainer (2014) proposed that these two quantities be formulated as a ratio, and referred to it as *VAR*. They further found that *VAR* could be well estimated from the reliability of the subscore and the disattenuated correlation between the subscore and the remainder score (total test without items from the subscore), where the disattenuated correlation is equal to the raw correlation divided by the square root of the product of subscore remainder score reliabilities. Let $r1$ be the reliability of the subscore and $r2$ be the disattenuated correlation between the subscore and remainder. Then,

$$VAR = 1.15 + 0.51r1 - 0.67r2 \quad (5)$$

Data Source

Data were obtained for a high-stakes assessment in a health profession completed by 539 examinees. The test consists of approximately 350 items partitioned into 7 subscores (e.g., anatomy, pathology), with each subtest consisting of 40 to 60 items (see Table 1 for details). Both *VAR* and *PVEV* were computed as well as other summary statistics.

Results

VAR

For each subtest, Table 1 presents means and SD on a proportion correct metric, subtest reliability, disattenuated correlation with remainder score, and *VAR*s. The total score mean was 0.64, with subtest means ranging from 0.52 to 0.69. Subscore coefficient alphas ranged from 0.54 to 0.78, and disattenuated correlations of the 7 subscores with remainder scores were generally high, falling between 0.75 and 0.98. *VAR*s ranged from 0.79 to 0.96. As none of the values exceeded 1.0, it is concluded that *subscores are not useful* for any of the subtests.

Table 1. Descriptive Statistics for Subscores

Subscore	*Mean*	*SD*	α	r_{dis}	*VAR*
A	0.52	0.14	0.78	0.88	0.96
B	0.64	0.11	0.70	0.88	0.92
C	0.60	0.13	0.74	0.95	0.89
D	0.66	0.12	0.76	0.86	0.96
E	0.67	0.10	0.69	0.98	0.85
F	0.63	0.94	0.66	0.90	0.92
G	0.69	0.84	0.54	0.75	0.79

Subscore Error and Profile Variability

The typical examinee had a total score mean of 0.64, with a conditional standard error of about 0.025. In contrast, the standard errors for the typical examinee on the seven subtests ranged from about 0.060 to 0.073, with the larger standard errors associated with less reliable subtests. Of course, the standard errors for individual examinees varied as a function of their number-right score. Profile SDs ranged from 0.028 to 0.199 with a mean of 0.090. The index of profile generalizability (Brennan, 2001) was also computed for the total group, with $\mathcal{G}$= .59.

***Figure 1*:** Mean conditional SEMs and score profile SDs at five ability levels (n = 539).

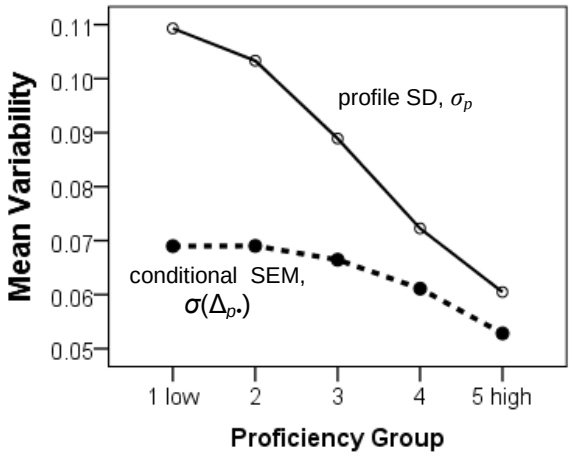


Figure 1 shows mean values of σ^2_p and $\sigma^2(\Delta_p.)$ for examinees in five ability groups. Note that computations use variances; however, graphs are based on the square roots of variances (i.e., standard deviations and standard errors) in order to put the relevant quantities onto the observed score scale. The graph indicates that score profiles generally were more variable than expected just based on random error, and variability was greatest for low proficiency examinees. Notably, the difference between the two measures is greatest for low scoring examinees.

Figure 2 shows the distribution of *PVEV* for the entire group of examinees. Values of *PVEV* ranged from 0.2 to 10.3, with a median of 1.8 and a mean of 2.0. Larger values of *PVEV* are more likely to be associated with score profiles that have meaningful variability. The key is to establish a threshold above which most of the score profiles will carry meaning. Clearly, 1.0 is too low, and even a value of 2.0 seems too low, as 2.0 would indicate that the observed score variability is only twice that expected due just to measurement error.

***Figure 2*:** Frequency Distribution for *PVEV* (n = 539)

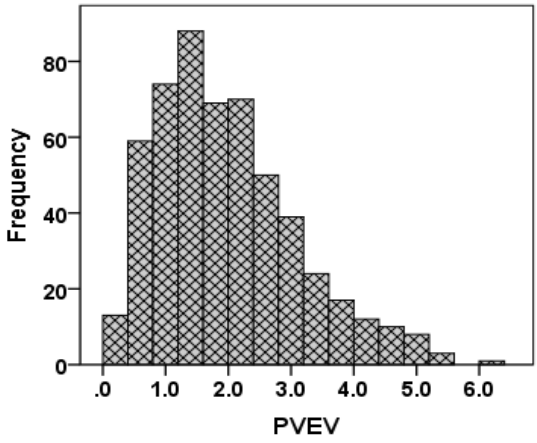
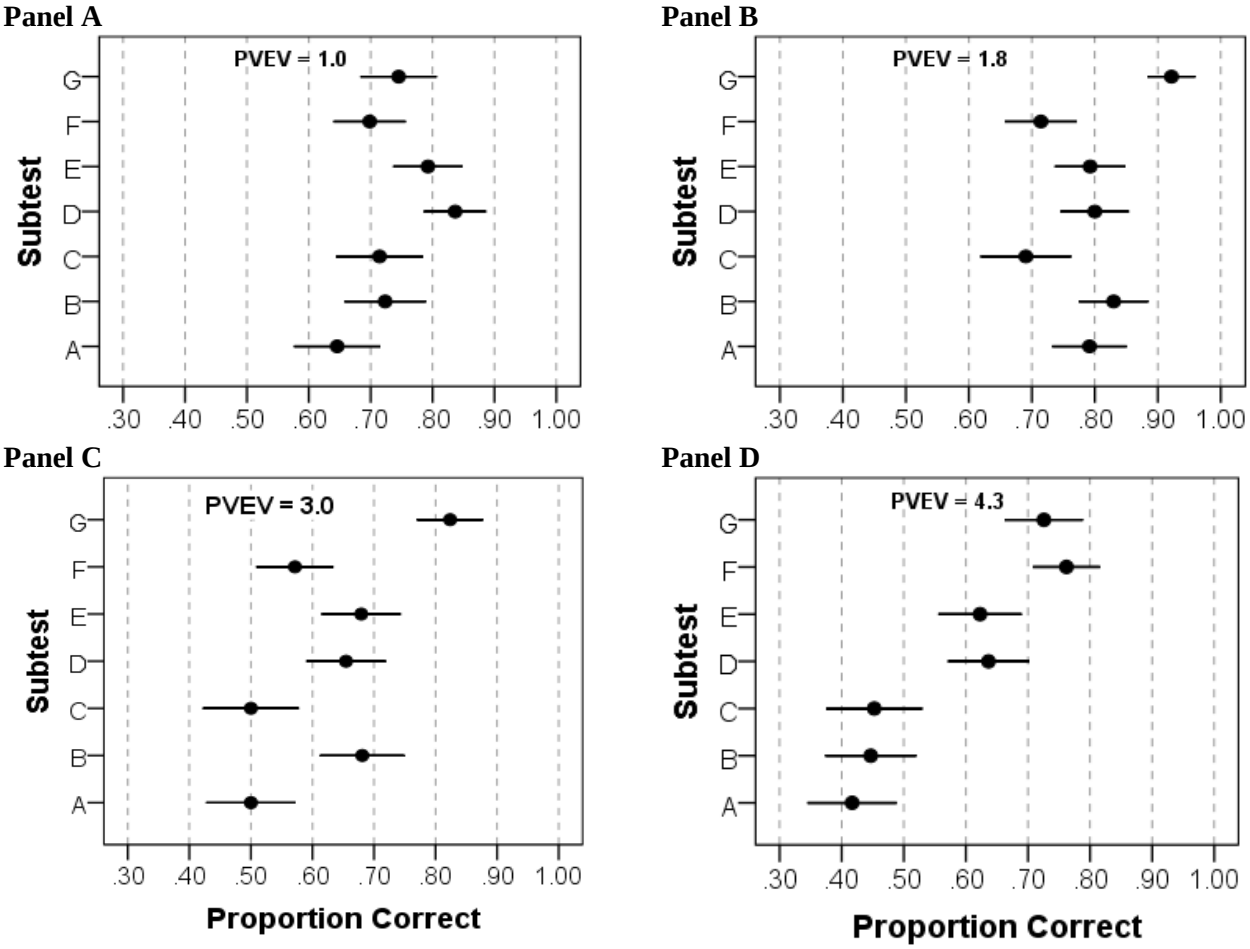


Figure 3 illustrates the degree of profile variability and conditional *SEM*s for score profiles at four values of *PVEV*: 1.0, 1.8, 3.0 and 4.3. These values correspond to the 44th, 50th, 83rd, and 95th percentiles of *PVEV*. The dot corresponds to the score for a person on each subtest, with the whiskers indicate ±1.0 *SEM*. Profiles for the examinees depicted in Panels A and B do not exhibit much variability, although subscore G in Panel B does appear to be meaningfully higher than other subscores, due in part to the small standard error associated with

high scores. Meanwhile, the profiles in Panels C and D appear to be quite variable. These two examinees might actually benefit by focusing their remedial activities on the content covered by certain subtests.

Figure 3: Illustrative Score Profiles at Different Levels of *PVEV*.



Discussion

Additional Observations

The goal of this paper is to demonstrate a different framework for evaluating the utility of subscores. We have analyzed other sets of test scores besides the data presented here and have three useful observations to add. First, this method provides a promising way to evaluate subscore utility for different subgroups. Although PRMSE recently has been suggested for such purposes (Sinharay & Haberman, 2014), group-splitting can wreak havoc on correlations and reliability coefficients if the splitting produces a restriction of range. Using data from a medical licensing test, we evaluated subscores for groups of examinees with different types of medical education (US vs international medical schools), and found that international students exhibited more variable profiles and larger values of *PVEV*. The finding is consistent with other reports that graduates of international schools have more variable curricular experiences.

Second, we have used *PVEV* to evaluate subscores based on alternate classification schemes for items from the same test. It is common to ask, "Does it make more sense to report scores based on content (e.g., disease categories) or on process (e.g., diagnosis, treatment)?" While such questions sometimes can be answered within the correlational and factor analytic traditions, the results are not always helpful. The methods

illustrated here may prove to be more sensitive to subtle differences among score profiles based on different frameworks. On a related note, we recently found that *PVEV*, when coupled with correlational analyses, provided very convincing evidence to policy makers to abandon certain subscore schemes.

Third, analyses of subscores from other tests indicate a very strong, but imperfect, empirical relationship between mean *PVEV* and $\mathcal{G}$. Although additional work is needed, we suspect two reasons why the relationship is not 1.0. First, to maintain consistency with other computations in multivariate generalizability theory, Brennan (2001) uses biased estimates of the variances that make up $\mathcal{G}$ (i.e., n rather than n-1), while computations presented here are unbiased. Second, $\mathcal{G}$ corresponds to *relative* error, *PVEV* in the present investigation is based on *absolute error*. It is possible to compute *PVEV* for conditional relative errors as well.

Limitations and Additional Research

While this paper advocates the use of *PVEV*, it does not offer guidelines regarding interpretation and use. Additional research into its relationship with conventional indices (e.g., *VAR*, $\mathcal{G}$) will help identify a threshold above which scores have meaningful variability. Given that $\mathcal{G}$ can be interpreted as a reliability-like index, it would be useful to determine the value of *PVEV* that correspond to a particular level of $\mathcal{G}$ that users of other reliability coefficients have grown accustomed to (e.g., .80). Another avenue would be to determine if the distribution of *PVEV* as an *F*-ratio can serve as a tool for interpreting *PVEV*.

It is important to recognize that the measures of profile variance and error variance proposed here are based on raw percent correct scores. It is common practice, albeit not always wise, to standardize scores or otherwise scale away differences in difficulty among subtest. Scaling all subtests to identical means will affect the results: it will reduce profile variability if raw subtest means differ. The impact on error variances is more complicated. Although it seems possible to transform error variances to any scale, such a transformation may cloud the interpretation of observed score conditional errors.

Another interpretative issue that needs to be worked out – for this or any context where subscores are reported – is the width of confidence intervals (CIs) and factors that affect their width. The error bars in Figure 3 are based on 1 *SEM*. One might argue that 90% or 95% CIs should be used. As with all statistical inference, one needs to weigh the costs of a Type I error, which will be greater in some instances than in others. Another factor that affects the CIs is whether an adjustment is to be made for multiple comparisons, which will increase the width of the CI. And, the choice of absolute versus relative *SEM*s also affects the CI. The present study was based on absolute errors which are almost always larger than relative errors. Using the relative errors would result in *smaller* CIs than reported in Figure 3. Furthermore, if score differences are subject to formal comparison (as when computing difference scores), then one also needs to consider the covariances between subtests, which will reduce width of the error band for difference scores.

Finally, there are similarities and difference between observed score conditional *SEM*s and indices derived from item response theory (IRT). One notable difference pertains to the relationship between conditional errors and total scores. Observed-score conditional *SEM*s are smaller toward the low and high end of the score distribution (inverted U), while conditional errors in IRT are larger at the ends of the theta distribution (U-shaped curve). The similarity between observed-score conditional *SEM*s in generalizability theory and person-fit indices in IRT is that both indicate the extent to which an examinee's item responses are consistent with the expected responses, given each item's difficulty and examinee's proficiency. Person fit in IRT seems more similar to the relative conditional *SEM* than to the absolute conditional *SEM* because the relative index includes a term corresponding to the covariance between an examinee's item responses and vector of item difficulties. The relationships between person fit and conditional *SEM*s is worthy of further exploration as both may signal the presence of multidimensionality.

Conclusions

Methods for evaluating subscores based on correlations and reliabilities are not sensitive to individual differences and run the risk of overlooking the possibility that some subscores will be useful for some people. Given that score profile variability and score precision vary by examinee ability and possibly by subgroup, it seems reasonable to suspect that the utility of subscores also may vary by examinee or subgroup. The present results demonstrate that there may be merit in reporting subscores for a small percentage of examinees even though conventional methods for evaluating subscores indicate that subscores should not be reported. Although the present findings are limited by the fact that they are based on subscores from a single test, they do support the investigation of new paradigms for evaluating subscore utility.

It is often assumed that if subscores do have value, it is because they can help examinees target their remediation efforts, usually by focusing on weak areas. However, as illustrated in Figure 3 (Panels B and C), it is the high scores that have greater precision and are therefore more interpretable. The smaller conditional *SEM*s for high scores suggest an obvious but often overlooked way to interpret subscores: rather than directing examinees toward what *to* study, subscores might be useful for indicating what *not to* study.

Finally, indices like *PVEV* may be more effective for encouraging dialogue between psychometricians and policy makers. Conventional statistical methods show that subscores are seldom meaningful, and yet subscores continue to be reported at the insistence of score users and policy makers. It would appear that policy makers have not heard the message being delivered by the psychometric press – in part because the message answers the wrong question. The method illustrated here is sufficiently liberal to encourage some subscore reporting, while highlighting instances where subscores will be futile. Such an approach may better engage policy makers in productive discussion by acknowledging that subscores may be meaningful for some examinees, but certainly not for everyone.

References

Brennan R.L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

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

Feldt, L.S., Steffen, M., & Gupta, N.C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement*, *9*, 351-361.

Haladyna, T.M., & Kramer, G. A., (2004). The validity of subscores for a credentialing examination. *Evaluation in the Health Professions*, *27*(4), 349-368.

Haberman, S.J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204-229.

Lord, F.M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, 15, 325-336.

Lord, F.M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement*, 17, 510-521.

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. (2013). A note on assessing the added value of subscores. *Educational Measurement: Issues and Practice*, *32*(4), 38-42.

Sinharay, S., & Haberman, S.J. (2014). An empirical investigation of population invariance in the value of subscores, *International Journal of Testing*, *14*:1, 22-48.