Article

Applied Psychological Measurement
2018, Vol. 42(8) 595–612
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621618758698
journals.sagepub.com/home/apm

The Use of Multivariate Generalizability Theory to Evaluate the Quality of Subscores

Zhehan Jiang[1] and Mark Raymond[2]

Abstract

Conventional methods for evaluating the utility of subscores rely on reliability and correlation coefficients. However, correlations can overlook a notable source of variability: variation in subtest means/difficulties. Brennan introduced a reliability index for score profiles based on multivariate generalizability theory, designated as $\mathcal{G}$, which is sensitive to variation in subtest difficulty. However, there has been little, if any, research evaluating the properties of this index. A series of simulation experiments, as well as analyses of real data, were conducted to investigate $\mathcal{G}$ under various conditions of subtest reliability, subtest correlations, and variability in subtest means. Three pilot studies evaluated $\mathcal{G}$ in the context of a single group of examinees. Results of the pilots indicated that $\mathcal{G}$ indices were typically low; across the 108 experimental conditions, $\mathcal{G}$ ranged from .23 to .86, with an overall mean of 0.63. The findings were consistent with previous research, indicating that subscores often do not have interpretive value. Importantly, there were many conditions for which the correlation-based method known as proportion reduction in mean-square error (PRMSE; Haberman, 2006) indicated that subscores were worth reporting, but for which values of $\mathcal{G}$ fell into the .50s, .60s, and .70s. The main study investigated $\mathcal{G}$ within the context of score profiles for examinee subgroups. Again, not only $\mathcal{G}$ indices were generally low, but it was also found that $\mathcal{G}$ can be sensitive to subgroup differences when PRMSE is not. Analyses of real data and subsequent discussion address how $\mathcal{G}$ can supplement PRMSE for characterizing the quality of subscores.

Keywords

subscores, score profiles, generalizability theory, reliability, dimensionality, simulation.

Examinees and other test users expect to receive subscores in addition to total test scores (Brennan, 2011; Huff & Goodman, 2007). As noted in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014), when the interpretation of subscores, score differences, or profiles is suggested, the testing agency is

[1] The University of Alabama, Tuscaloosa, AL, USA
[2] National Board of Medical Examiners, Philadelphia, PA, USA

Corresponding Author:
Zhehan Jiang, Assistant Professor, The University of Alabama, 309E Gorgas Library, Tuscaloosa, AL 35487, USA.
Email: zjiang17@ua.edu

obligated to demonstrate that subscores exhibit sufficient reliability and are empirically distinct. Simple correlations have long been used to evaluate the distinctiveness of subscores (e.g., Haladyna & Kramer, 2004), as have factor analytic methods (Stone, Ye, Zhu, & Lane, 2010; Thissen, Wainer, & Wang, 1994). More recently, a method developed by Haberman (2008) incorporates both subscore distinctiveness and subscore reliability into a single decision rule about whether to report subscores. The idea of the method is that an observed subscore, S, is meaningful only if it can predict the true subscore, S_T, more accurately than the true subscore can be predicted from the total score Z, where S_T is estimated using Kelley's equation for regressing observed scores toward the group mean, and where predictive accuracy is expressed as mean-square error. If the proportion reduction in mean-square error (PRMSE) based on the prediction of S_T from S exceeds the PRMSE based on the total score Z, then the subscore adds value—that is, observed subscores predict true subscores more accurately than total scores predict true subscores. Feinberg and Wainer (2014) suggested that the two PRMSE quantities be formulated as a ratio, and referred to it as the value-added ratio (VAR). If VAR exceeds 1, then the subscore is deemed useful. In a related vein, Brennan (2011) proposed a utility index, U, which produces the same decisions as PRMSE and VAR but is computationally more straightforward than PRMSE. This article uses VAR to refer to PRMSE, U, and VAR.

A consistent finding from numerous studies of operational testing programs is that subscores are seldom worth reporting (Puhan, Sinharay, Haberman, & Larkin, 2010; Sinharay, 2010, 2013; Stone et al., 2010). Although well-constructed test batteries used for selection and admissions can produce useful subscores for their major sections (e.g., reading, math), the subscores reported within the major sections often lack empirical support (Haberman, 2008; Harris & Hanson, 1991). The overreporting of subscores is particularly prevalent on licensure and certification tests (Puhan et al., 2010; Sinharay, 2010).

Although VAR provides a useful framework for evaluating subscores, there are conditions where additional information would help inform decisions about reporting subscores: First, VAR indicates whether subscores are reliably different from the total score but does not address other properties of subscores. Test users are sometimes interested in questions such as ''Are these two subscores different from each other?'' Or, ''To what extent does my score overall profile reliably indicate my strength and weaknesses?'' Second, Feinberg and Jurich (2017) noted that the criterion of VAR $>$ 1.0 is often too liberal, and encourages reporting subscores that do not explain a meaningful amount of variance above the total score. Third, correlation-based methods such as VAR are not sensitive to systematic differences in subtest difficulty, and consequently can overlook differences in score profiles for subgroups of examinees (Cronbach & Gleser, 1953; Livingston, 2015). Consider, for example, two subscores, X and Y, with equal means ($M_X = M_Y = 0$, and a correlation, $r_{xy} = .90$). Assume that one half of examinees in a large sample are exposed to an intervention that results in a score increase of 0.50 *SD* to the scores on Y. The new correlation for the total group would be $r_{xy} = .873$, a change of only .027 from the original value. This contrived example has a real-world counterpart involving gender differences on essay tests. It has been documented that females generally score higher than males on essay questions and lower on multiple-choice questions (MCQs); and, although the gender differences in score profiles are considerable (Bridgeman & Lewis, 1994; Sinharay & Haberman, 2014), correlational methods suggest that essay scores should be combined with multiple-choice scores rather than separately reported (e.g., Bridgeman & Lewis, 1994; Thissen et al., 1994). Not reporting subscores in this instance could result in missing potentially important performance differences between men and women (Bridgeman & Lewis, 1994), or overlooking the incremental validity of essay scores for predicting certain outcomes (Bridgeman, 2016). In short, correlational methods focus primarily on between-subject variation, and there are circumstances for which it is useful to also consider within-subject variation.

Subscore Profile Reliability and Multivariate Generalizability Theory

Score profiles are what testing agencies report and what test users interpret; thus, it seems natural to inspect the properties of actual score profiles when evaluating the utility of subscores. This amounts to the study of within-subject variation (Brennan, 2001; van der Maas, Molenaar, Maris, Kievit & Borsboom, 2011). Flat score profiles can be said to contain no information above that provided by the total score, whereas variable profiles may contain additional information. The challenge is to differentiate signal from noise, or true score variance from error variance, in score profiles. Cronbach, Gleser, Nanda, and Rajaratnam (1972) laid the groundwork for differentiating signal from noise in score profiles within the context of multivariate generalizability theory (G-theory). Their efforts remained partially developed and obscure until integrated into a comprehensive framework by Brennan. Brennan (2001) introduced a reliability-like index for score profiles designated as $\mathcal{G}$, which *indicates the proportion of variance in observed score profile variance attributable to universe (or true) score profile variance* (Brennan, 2001). One important difference between $\mathcal{G}$ and PRMSE is that $\mathcal{G}$ treats the profile as the unit of analysis. That is, $\mathcal{G}$ characterizes the entire score profiles rather than each specific subtest.

The G-theory design most relevant to the study of subscores is where a different set of items (i) is assigned to each of multiple subtests (v), and all persons (p) respond to all items within each subtest. The univariate designation for this design is persons crossed with items nested within subtests, or p x (i:v). The multivariate designation of this design is $p\bullet$ x i°, where the circles describe the multivariate design. In this instance, there is a random-effects p x i design for each level of some fixed facet. The solid circle indicates that every level of the person facet is linked to each level of the multivariate facet (i.e., linked with each subtest), whereas the open circle indicates that items are not linked across the different subtests (i.e., each subtest comprises a unique set of items).

A multivariate G study based on the $p\bullet$ x i° design produces matrices of variance–covariance components for persons, items, and error, designated as $\mathbf{\Sigma}_p$, $\mathbf{\Sigma}_i$, and $\mathbf{\Sigma}_\delta$. Also of interest is **S**, the observed variance–covariance matrix. **S** is equal to the sum of the variance–covariance component matrices $\mathbf{\Sigma}_p$ and $\mathbf{\Sigma}_\delta$; alternatively, it can be computed directly from observed scores. Brennan (2001) defined the generalizability index for score profiles as

$$\mathcal{G} = \frac{\mathcal{V}\left(\mu_p\right)}{\mathcal{V}(\bar{X}_p)} = \frac{\left[\overline{\sigma_v^2}(p) - \overline{\sigma_{vv'}}(p)\right] + var(\mu_v)}{\left[\overline{S_v^2}(p) - \overline{S_{vv'}}(p)\right] + var(\bar{X}_v)}, \tag{1}$$

where $\mathcal{V}(\mu_p)$ is the average variance of universe score profiles, and $\mathcal{V}(\bar{X}_p)$ corresponds to the average variance for observed score profiles. $\mathcal{G}$ ranges from 0 to 1, and can be interpreted as a reliability-like index for score profiles. The numerator includes the following:

$\overline{\sigma_v^2}(p)$ = mean of the universe score variances for n_v subtests, given by the diagonal elements in $\mathbf{\Sigma}_p$;

$\overline{\sigma_{vv'}}(p)$ = mean of the all n_v elements in $\mathbf{\Sigma}_p$; and

$var(\mu_v)$ = variance of the subscore means, which is estimated by $var(\bar{X}_v)$.

Meanwhile, the denominator is defined as

$\overline{S_v^2}(p)$ = mean of the observed score variances obtained from the diagonal elements in **S**;

$\overline{S_{vv'}}(p)$ = mean of all n_v elements in **S**;

$var(\bar{X}_v)$ = variance of the subscore means.

One convenience is that $var(\bar{X}_v)$ provides an estimate of $var(\mu_v)$. Another is that for the $p\bullet$ x i° design, the covariance components for observed scores provide an unbiased estimate of covariance components for universe scores. That is, $\sigma_{vv'} = S_{vv'}$, or the off-diagonal elements of $\mathbf{\Sigma}_p$ equal the off-diagonal elements of $\mathbf{S}$.

Equation 1 has a few noteworthy features when applied to the $p^\bullet \ x \ i^\circ$ design: First, ignoring for the moment the right-hand terms (i.e., the variance of the means), it can be seen that $\mathcal{G}$ is essentially the average of the subtest reliabilities adjusted downward for the subtest covariances. Specifically, the left-most terms in the numerator and denominator, $\overline{\sigma_v^2}(p) \ / \ \overline{S_v^2}(p)$, represent the ratio of average true score variance to average observed score variance, which is essentially the average of subtest reliabilities. Also, covariances are subtracted out of both the numerator and denominator. Thus, as subscore correlations approach 1, the value of $\overline{\sigma_v^2}(p) - \overline{\sigma_{vv'}}(p)$ decreases, as does $\overline{S_v^2}(p) - \overline{S_{vv'}}(p)$. As $\overline{\sigma_v^2}(p)$ is almost always less than $\overline{S_v^2}(p)$, the quantity $\overline{\sigma_v^2}(p) - \overline{\sigma_{vv'}}(p)$ declines more quickly, resulting in a decrease in $\mathcal{G}$. Second, any differences in subtest means will contribute positively to $\mathcal{G}$, as long as $\left[[\overline{\sigma_v^2}(p) - \overline{\sigma_{vv'}}(p) \right]$ is less than $\left[\overline{S_v^2}(p) - \overline{S_{vv'}}(p) \right]$, which is typically the case. By extension, if score profiles are flat for one group ($var(\bar{X}_v) = 0$) but variable for a second group ($var(\bar{X}_v) > 0$), then $\mathcal{G}$ will be higher for the latter group all other things being equal. Third, it is evident that if subtests correlate 0 and subtest means are equal, $\mathcal{G}$ equals the average of the subtest reliabilities. Subtest reliability places an upper limit on $\mathcal{G}$ when means are equal. Although it appears as if $\mathcal{G}$ can exceed subtest reliability, it likely would require low subtest correlations and considerable variance in subtest means.

Brennan (2001) provided an example where $\mathcal{G}$ is computed for three subtests from the mathematics section of a college entrance test. Each subtest contains about 20 items, with reliability coefficients of .78, .79, and .82. The disattenuated correlations among the three subtests are in the low .90s. While the subtests are sufficiently reliable for subscores, the correlations suggest that the scores are not very distinct. As it turns out, $\mathcal{G} = .57$, indicating that 57% of the variance in observed score profile variance can be attributable to true score profile variance. Of course, the question is whether a $\mathcal{G}$ of .57 is sufficiently high to support reporting subscores. To date, $\mathcal{G}$ has received little attention in the literature, and there is no practical guidance regarding its interpretation.

The purpose of this article was to report the results of a series of studies undertaken to investigate the properties of $\mathcal{G}$ under various conditions of practical interest. The first study consists of three pilot experiments that use simulated item responses to evaluate $\mathcal{G}$ within the context of a single examinee group taking a certification test. Given the lack of prior research on $\mathcal{G}$, the primary objective of this pilot effort was to learn more about its general behavior under a broad range of testing conditions. The second simulation experiment is the main focus of this article; it evaluates the sensitivity of $\mathcal{G}$ to subgroup differences in score profiles. More specifically, the goal is to determine the extent to which $\mathcal{G}$ detects differences in the reliability of score profiles for subgroups of examinees when the groups have similar covariance structures but different subscore means, which is a common occurrence in both educational and certification testing (e.g., Sinharay & Haberman, 2014). Finally, to illustrate how $\mathcal{G}$ might be interpreted in practice, a third study computes $\mathcal{G}$ for subscores obtained from an actual certification test for which the testing agency hypothesized that subscores may have more utility for one subgroup of examinees. To provide a context for interpreting $\mathcal{G}$, VAR is also computed for each study. The intent is not to so much compare $\mathcal{G}$ and VAR but rather to use familiarity with VAR to better understand the potential utility of $\mathcal{G}$.

Pilot Experiment: Total Group $\mathcal{G}$

The appendix contains the complete methods and results for the pilot study; a summary is provided here. Three experiments evaluated the response of $\mathcal{G}$ to different conditions of subtest reliability (ρ_v^2), subtest correlations ($\overline{\rho_{vv'}}$), and variation in subtest means ($var(\mu_v)$). Each study consisted of three levels of $\overline{\rho_{vv'}}$, three levels of $var(\mu_v)$, and four levels of ρ_v^2, for a total of 36 conditions per study. Study A simulated a total test score partitioned into two highly reliable subtests with the four levels of ρ_v^2 ranging from .85 to .88. Study B consisted of four moderately reliable subtests with ρ_v^2 ranging from .77 to .83. Study C consisted of six less reliable subtests, with ρ_v^2 ranging from .66 to .78. For all three studies, the three levels of $\overline{\rho_{vv'}}$ = .70, .80, and .90, and the levels of $var(\mu_v)$ = .06, .25, and .56. About 120 replications were run for each of the 36 × 3 = 108 conditions, with N = 1,000 simulated examinees per replication. Both $\mathcal{G}$ and VAR were computed for each replication and averaged across replications within a condition.

As expected, $\mathcal{G}$ increased with higher levels of subtest reliability, greater differences in subtest means, and lower levels of subtest correlation. These main effects accounted for about 95% of the variation in observed means across the three studies. The only notable interaction effect occurred between $\overline{\rho_{vv'}}$ and $var(\mu_v)$, with greater variation in subtest means diminishing the impact of subtest correlations on $\mathcal{G}$. A less expected finding was that $\mathcal{G}$ seldom reached conventionally acceptable levels of reliability. Across all 108 conditions, $\mathcal{G}$ ranged from .23 to .86, with an overall mean of 0.63. Although VAR and $\mathcal{G}$ covaried, there were exceptions to the trend, in that conditions existed where VAR seemed quite generous (VAR $\geq$ 1), but $\mathcal{G}$ was low ($\mathcal{G} < .70$).

Main Experiment: $\mathcal{G}$ for Two Groups

Method

Design. The primary objective of Experiment 2 was to determine the extent to which $\mathcal{G}$ and VAR are differentially sensitive to differences in subtest means, reliabilities, and correlations for subgroups of examinees. In addition, the relationship was evaluated between $\mathcal{G}$ and VAR under more informative conditions. As with Experiment 1, conditions were created by manipulating subtest reliability, subtest correlation, and variability in subscore means. For each condition, there was a *reference group* whose score profile was flat, and a lower performing *focal group* whose score profile varied. It was hypothesized that $\mathcal{G}$ would be higher for the focal group due primarily to the differences between subscore means for that group. The use of two groups corresponds to circumstances common in practice (e.g., male vs. female; minority vs. nonminority; English primary vs. English language secondary), and will simplify interpretation. The results based on two groups are expected to generalize to three or more groups.

Data were simulated for instances in which there are two subtests. The pilot experiment and work by Sinharay (2010) suggest that results for two subtests generalize to multiple subtests. Also, the two subtest case corresponds to a common data interpretation challenge in testing: whether to report subscores when a test consists of MCQs and some other format such as constructed responses (Bridgeman, 2016; Bridgeman & Lewis, 1994; Thissen et al., 1994). Within each group (reference and focal), three factors were investigated:

- Population (disattenuated) correlation, $\rho_{vv'}$, between subtests. Three levels were studied, with values of $\rho_{vv'}$ = .73, .81, and .90. These values are comparable with the subtest correlations often seen in the literature (e.g., Sinharay, 2010; Sinharay & Haberman, 2014).
- Subtest reliability, ρ_v^2. Three levels were studied, designated as high (ρ_v^2 = .89), moderate (ρ_v^2 = .83), and low (ρ_v^2 = .71). The two subtests were fixed to have equal reliabilities.
- Difference in the two subtest means, $\Delta\mu$, for the focal group. The four levels of $\Delta\mu$ for the focal group were set at .00, .25, .50, and .75. For example, the means in the first condition were 0.00 for both subtests (i.e., $\Delta\mu$ = .0), while the subtest means in the second condition were set at 0.00 and 0.25 (i.e., $\Delta\mu$ = .25). Meanwhile, the higher performing

reference group always had no variation in means, with $\bar{\mu}$ = 1.0 for both subtests. Although $\Delta\mu$ = .75 seems high, this level of variability in subscore profiles can occur in practice (e.g., Sinharay & Haberman, 2014).

These three factors were fully crossed, creating 36 conditions within each group for a total of 72 conditions. Sample size was set at N = 1,000 per examinee group for each of 120 replications per condition.

Data simulation. Item responses were simulated in essentially the same manner as described for the pilot study as documented in the appendix. Rather than using parameters from an actual certification test as was done for the pilot study, discrimination parameters were generated from a log-normal distribution (M = 0.0, SD = 0.5), while difficulty parameters were normally distributed (M = 0, SD = 1), which is common for simulation studies of this nature. True ability parameters for examinees were assumed to follow a multivariate normal distribution whose mean vector is $\boldsymbol{\mu}$ and covariance matrix is $\boldsymbol{\Sigma}_p$, where both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_p$ contained only two elements. The vectors of ability parameters $\boldsymbol{\mu}$ were specified for the focal group to produce to the values of $\Delta\mu$ described above and as presented in Figure 1. The diagonal elements of $\boldsymbol{\Sigma}_p$ were constrained to be 1 (i.e., correlation matrix). The off-diagonal value is designated as $\rho_{vv'}$ and was assigned values of .73, .81, and .90.

Outcome variables. Both $\mathcal{G}$ and VAR were computed for each replication. For each of the 36 conditions within each group, the mean $\mathcal{G}$ was reported across the 120 replications. The proportion of the 120 replications was reported for which VAR $\geq$ 1. Scatterplots were produced to evaluate the relationship between $\mathcal{G}$ and VAR.

Results

Figure 1 summarizes results for $\mathcal{G}$ on the left and VAR on the right. Vertically, each panel corresponds to a different level of subtest reliability (low = .71, moderate = .83, high = .89), while the lines within panels correspond to the three levels of subtest correlation ($\rho_{vv'}$ = .73, .81, .90). For the reference group, $\Delta\mu$ was always equal to 0, thus the x axis displays only one level for that group. Across all conditions, $\mathcal{G}$ ranged from .27 to .82, with an overall mean of 0.53. However, values of $\mathcal{G}$ for the reference group were consistently about .10 lower than those for the focal group. Within each panel in the left portion of Figure 1, $\mathcal{G}$ increased as the subtest correlations decreased and as variation in subtest means increased. Looking down the panels, it can be seen that $\mathcal{G}$ declines with lower levels of subtest reliability.

The values of $\mathcal{G}$ are generally modest, even for conditions where one might expect it to be high. For example, under the conditions most favorable to subscores (top panel, top line), $\mathcal{G}$ is only .70 for the reference group and ranged from .72 to .82 for the focal group. Under the least favorable conditions (bottom panel, bottom line), $\mathcal{G}$ was only .28 for the reference group and ranged from .29 to .52 for the focal group. A key finding is that in all conditions where focal group score profiles were not flat (i.e., $\Delta\mu > 0$), their $\mathcal{G}$ index was higher than that for the reference group; these differences can be attributed primarily to the sensitivity of $\mathcal{G}$ to variation in subtest means for the focal group.

The panels on the right side of Figure 1 show the proportion of replications for which VAR was greater than 1.0. The mean VARs across all conditions and all three panels were 0.46 for the reference group and 0.50 for the focal group, indicating some sensitivity of VAR to the differences in focal group score distributions. As Figure 2 implies, the distribution of VAR was bimodal: For about half the conditions, subscores were worth reporting, and for half they were not. Results indicate that subscores are usually worth reporting for moderate to high levels of

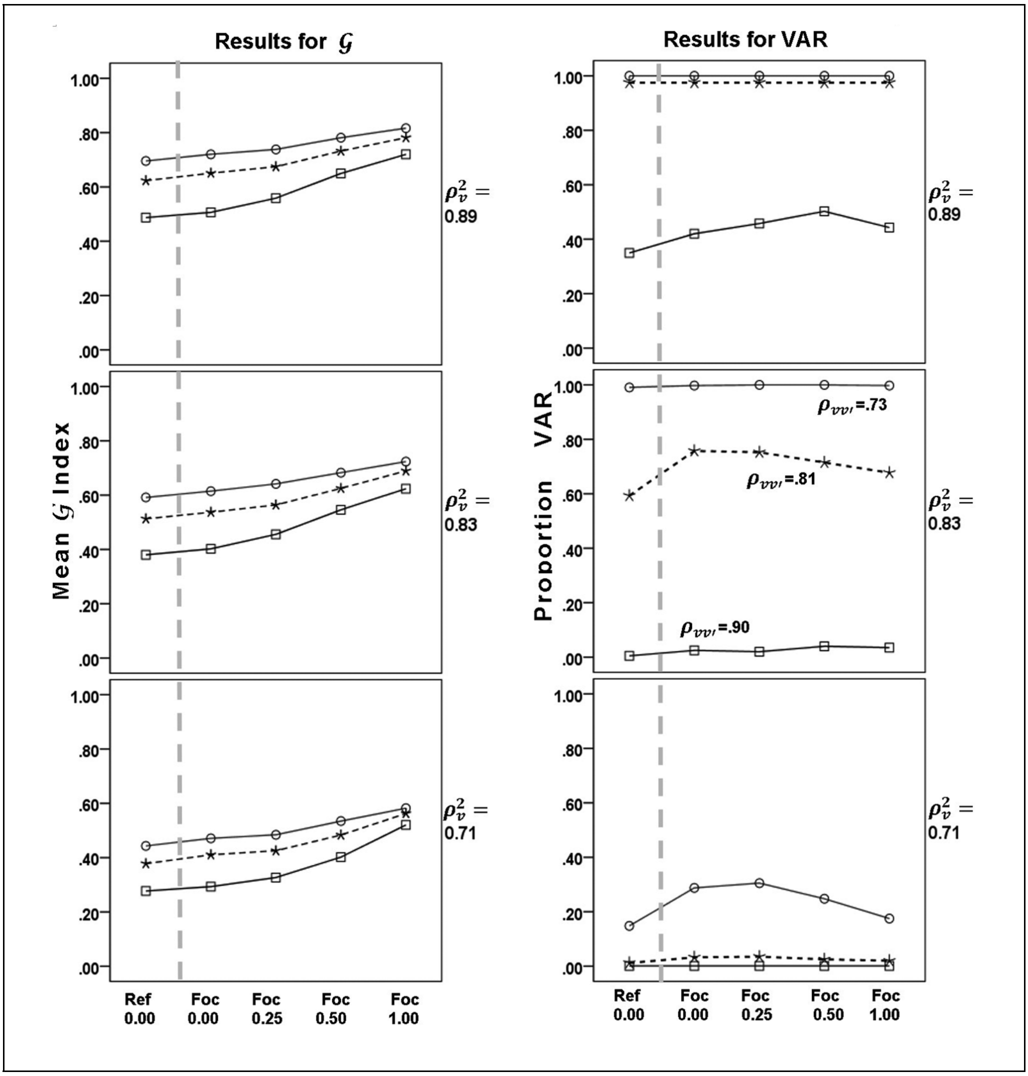

Figure 1. Mean $\mathcal{G}$ and proportion VAR as a function of differences in subtest means ($\Delta\mu$) for the reference (Ref) and focal (Foc) groups.

Note. There are three levels of subtest reliability (ρ_v^2) arranged vertically and levels of true score correlation ($\rho_{vv'}$) within each panel. The reference subtest means were always equal ($\Delta\mu$ = .00). VAR = value-added ratio.

reliability (ρ_v^2 = .83, .89), and when subtest correlations are not excessive ($\rho_{vv'}$ = .73, .81). These findings are not unlike those reported by Sinharay (2010). On occasion, VAR supported reporting subscores for the focal group more often than the reference group (ρ_v^2 = .83 and $\rho_{vv'}$ = .91). This is a consequence of the focal groups' score distributions being more variable than the distribution for the reference group, resulting in higher subtest reliabilities, which in turn improves VAR. This outcome is consistent with the results of Study 1 (appendix) and with Brennan (2011), which demonstrates that VAR-like indices are particularly sensitive to small changes in reliability when the subtest–total test correlations are high.

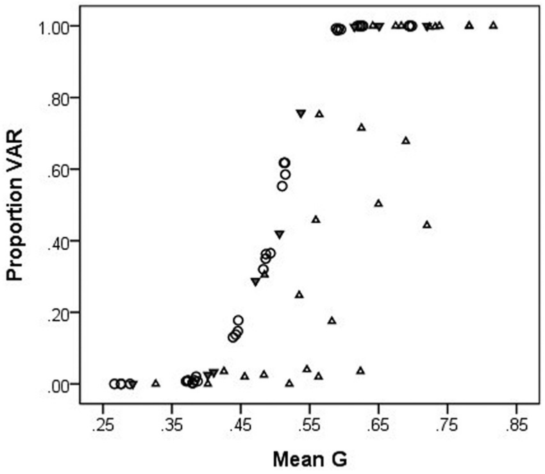

Figure 2. Relationship between proportion VAR and mean $\mathcal{G}$ under all experimental conditions.
Note. Open circles indicate the reference group ($\Delta\mu$ = 0); inverted triangles indicate the focal group where $\Delta\mu$ = 0; while the upright triangles indicate the focal group where $\Delta\mu > 0$. If data points where $\Delta\mu > 0$ are excluded, the relationship between mean $\mathcal{G}$ and proportion VAR can be modeled with a logistic function (R^2 = .95). VAR = value-added ratio.

Figure 2 presents a scatterplot of $\mathcal{G}$ with VAR. The open circles correspond to the reference group, whereas the triangles correspond to the focal group. The inverted triangles indicate that $\Delta\mu$ = 0 for the focal group, whereas the upright triangles indicate $\Delta\mu > 0$. One notable observation is that for all conditions where $\Delta\mu$ = 0, which includes all 36 reference group conditions and nine focal group conditions, VAR indicates that subscores are worth reporting most of the time when values of $\mathcal{G}$ exceed about .55. In other words, VAR greater than 1.0 does not guarantee that the score profiles will be very reliable. A second observation is that for all conditions where $\Delta\mu$ = 0, the relationship between proportion VAR and mean $\mathcal{G}$ follows a fairly tight S function. By replacing a few zeros for VAR with near-zero values, a logistic model with R^2 = .95 could be fit when $\Delta\mu$ = 0. A third point illustrated in Figure 2 is that for focal group conditions where $\Delta\mu > 0$, the two indices diverge from their monotonic relationship, and $\mathcal{G}$ looks better than what is expected based on VAR.

Example With Real Test Scores

Data Source

Item responses were available for a certification test in a health profession administered to 3,491 examinees. The 200-item test comprises five subtests for which reliabilities range from .73 to .83. The first four subtests consist of technical content (e.g., radiation biology; medical imaging), while the fifth subtest addresses human interactions with patients and other staff (e.g., communication; ethics). The reference group consists of 2,399 traditional examinees who recently completed a formal educational program and were seeking initial certification, whereas the focal group consisted of 92 examinees who had completed their education several years earlier, had let their certification lapse, and were seeking recertification. The testing agency thought that subscores might be particularly beneficial for the focal group because those examinees were

Table 1. Means (*M*), Standard Deviations (*SD*), Subtest Reliabilities (ρ^2_v), Subtest–Total Correlations ($\rho_{v,z}$), and VAR for Reference and Focal Groups on Five Subtests.

	M (% correct)		*SD* (% correct)		(ρ^2_v)		$(\rho^{)}_{v,z}$		VAR	
Subtest	Reference	Focal	Reference	Focal	Reference	Focal	Reference	Focal	Reference	Focal
A	0.78	0.66	0.12	0.16	.77	.84	.87	.84	.89	1.13
B	0.72	0.60	0.17	0.18	.75	.73	.85	.81	.86	.91
C	0.75	0.66	0.14	0.17	.82	.81	.91	.90	.90	.90
D	0.79	0.67	0.13	0.13	.83	.79	.89	.87	.94	.97
E	0.83	0.82	0.11	0.12	.73	.79	.75	.75	1.07	1.24

Note. The $\mathcal{G}$ for the reference group and focal group were .51 and .71, respectively. VAR = value-added ratio.

less likely to have received much recent feedback regarding their skills. The question of interest is whether subscore profiles for either of the groups are sufficiently reliable to report. For each group, descriptive statistics and VAR were obtained for the five subtests, and $\mathcal{G}$ was obtained for the score profile.

Results

Table 1 presents the relevant statistics for each examinee group. Test scores as reported here are on a percent correct metric. The column of means indicates that the score profiles for the two groups are nearly parallel, with the focal group obtaining consistently lower scores on four of the five subtests, and the two groups having equal scores on the fifth subtest. The score differences on Subtests A through D are not surprising given their technical content, while the similarity for Subtest E is consistent with the very practical skills covered by that subtest. Note also that the focal group exhibits greater variability on most subtests.

Table 1 also presents reliability coefficients (ρ^2_v) for each subtest, and the observed correlations between each subscore and the total score $(\rho^{)}_{v,z}$. Subtests A and E were more reliable for the focal group than for the reference group, whereas Subtests B, C, and D are about equally reliable for the two groups. Subtest–total score correlations for the two groups were similar, differing by a maximum of .04 for Subtest B. Overall, Table 1 suggests that the two groups are very similar except for the difference in means and a couple of the reliability coefficients.

Four of the five VAR indices for the reference group are less than 1.0; the exception is Subtest E, suggesting that it may add value above the total score. The VAR indices for the focal group were higher than those for the reference group. Subtests A and E were deemed worth reporting with indices of 1.13 and 1.24. The higher VAR for the focal group for Subtest A can be attributed to its higher reliability (.77 vs. .84) and its lower correlation with total score (.87 vs. .84). The higher focal group VAR for Subtest E is explained completely by the higher reliability of subscores for Subtest E. The $\mathcal{G}$ index for reference group score profiles was .51, indicating that only about half the variability in the typical profile could be attributed to true score variance. Meanwhile, subscore profiles for the focal group were considerably more reliable, with $\mathcal{G} = .71$.

Given that VAR exceeded 1.0 for two of the subtests for the focal group, and that $\mathcal{G}$ was a modestly respectable .71, a testing agency might decide to report the entire score profile for that group. Alternatively, the agency might decide to report Subscores A and E as they are but combine Subscores B, C, and D into a single score. Such a decision would produce an overall $\mathcal{G} = .83$ for the focal group. The VAR indices for the three subscores would be 1.13, 0.99, and 1.24.

Reporting the same three subtests for the reference group would result in $\mathcal{G}$ = .57, while the VAR indices for those three subscores would be .89, .99, and 1.07. The higher values of VAR for the focal group can be attributed in part to the increased heterogeneity of the focal group, which positively affected focal group reliability for Subtests A and E. The difference in $\mathcal{G}$ coefficients between the two groups is notable, and makes it difficult to overlook that score profiles are more reliable for the focal group. If a testing agency were to strictly follow VAR guidelines, it would report two subscores for the focal group and one for the reference group. However, the values of $\mathcal{G}$ and a desire for simplicity in reporting policy might suggest reporting three subscores for the focal group and none for the reference group. This example illustrates how both VAR and $\mathcal{G}$ can be used jointly, by first applying VAR to eliminate or combine subscores, and then computing $\mathcal{G}$ to document the reliability of the entire profile.

Discussion

This initial investigation of $\mathcal{G}$ shed some light on its potential utility for evaluating the quality of subscores. Consistent with expectations, the simulation experiments indicated that $\mathcal{G}$ increases with higher levels of subtest reliability, lower subtest correlations, and greater differences in subtest means. In addition, large differences in subtest means were found to attenuate the impact of subtest correlations on $\mathcal{G}$. One notable result was that $\mathcal{G}$ reached conventionally acceptable levels of reliability only under the most favorable conditions—when subtest reliabilities reached .80s, and subtest correlations were at .70 or .80. These conditions are not met very often in practice, where it is more typical to see subtest reliabilities in the .70s, and correlations in the mid-.80s and higher (Sinharay, 2010). Under these more typical conditions, $\mathcal{G}$ fell into and below the .60s, which is discouraging if one is looking for evidence to support subscores. Another notable finding was that there were numerous conditions for which VAR indicated that subscores should be reported even though $\mathcal{G}$ indices only reached the .50s and .60s. That is, VAR seemed quite tolerant of imprecise score profiles. However, for those conditions where score subtest means exhibited variability, $\mathcal{G}$ was more likely than VAR to suggest that subscores might be useful.

The main experiment provided additional insight into the relationship between $\mathcal{G}$ and VAR. When subscore profiles for the reference group and focal group were flat, the relationship between VAR and $\mathcal{G}$ could be accurately modeled by a logistic function (R^2 = .95); however, as score profiles varied ($\Delta\mu > 0$), $\mathcal{G}$ increased relative to VAR and their relationship weakened. The ability of $\mathcal{G}$ to detect group differences in subscore utility could be a useful area for further inquiry, particularly in those instances where it is important to acknowledge and understand subgroup differences based on gender, ethnicity, or language differences (e.g., Bridgeman & Lewis, 1994; Livingston, 2015).

Although they are related, VAR and $\mathcal{G}$ can lead to different conclusions as Figure 1 illustrated. Assume that a $\mathcal{G} > .70$ has been established as the threshold required for reporting subscores. Now consider the top-left and top-right panels of Figure 2 where subtests are very reliable (ρ_v^2 = .89), and the two lines where subtest correlations are less not extremely high ($\rho_{vv'}$ = .73, .81). Note that VAR would deem subscores to be worth reporting for *all* of these conditions for both the reference group and the focal group. However, adopting the guideline that $\mathcal{G}$ must exceed .70 would suggest that *none* of these conditions would produce reportable subscores for the reference group, and that subscores for the focal group would be reportable for six of eight conditions. In these conditions, VAR appeared to be somewhat lenient and less sensitive than $\mathcal{G}$.

Just as there are general guidelines regarding the interpretation of reliability coefficients and VAR, it is tempting to propose guidelines for $\mathcal{G}$. However, this single study is not sufficient to suggest minimum values of $\mathcal{G}$ or other specific recommendations for interpretation; such guidelines will develop as psychometricians gain experience with real and simulated data, much in the way that guidelines evolved for coefficient alpha. Instead, some general recommendations are offered for interpreting and applying $\mathcal{G}$: First, it must be recognized that $\mathcal{G}$ characterizes subscores for a *population* (or subpopulation) not for individual examinees. This is also true of coefficient alpha, subtest–total test correlations, and of VAR-related indices. Such indices inform testing agencies and institutional test users about the general properties of a test and the scores it produces, and are useful for deciding whether a test might be useful for some specific purpose. $\mathcal{G}$ should have a similar role for test batteries and tests for which subscores are reported. It should be useful for comparing two or more tests, or for monitoring tests and test scores over time. For example, a college admissions office might use $\mathcal{G}$ to determine if the subscores on College Entrance Exam A are more effective than subscores on Entrance Exam B for selecting some population of students into different majors. While $\mathcal{G}$ can support general statements about the usefulness of subscores for groups of examinees, it is not helpful for determining if an individual score profile is useful for decision-making purposes (e.g., choosing a major; deciding what content to study after failing a test). The same is true of VAR. Even if such indices indicate that subscores are useful for the group, there likely will be individuals for whom subscores are not informative. That is, $\mathcal{G}$ and VAR can tell a testing agency whether subscores are generally worth reporting but not whether the subscores within any particular profile are likely to be the same or different.

Making inferences about individuals requires evaluating the score profiles of individual examinees (Brennan, 2001). It is common practice to provide test users (e.g., admissions officers; teachers; examinees) with confidence intervals (CIs) around each subscore within an individual's score profile. One rule of thumb given to test users is that if the CIs for two subscores overlap, then the level of performance on the two subtests is probably the same. These CIs are often based on a group-level reliability index for that subtest, which means that their width for a particular subtest will be the same for all examinees. However, given that measurement error differs for individual examinees (e.g., AERA et al., 2014), CIs are more accurate when based on conditional standard errors of measurement (SEMs). Brennan (2001) provided equations necessary for computing observed score conditional SEMs within the multivariate G-Theory framework.

A second recommendation is based on the recognition that VAR and $\mathcal{G}$ have different purposes. VAR indicates if a subscore adds value above and beyond the information already available in the total score, while $\mathcal{G}$ characterizes the quality of an entire score profile. These differences suggest complementary roles for the two indices. Specifically, VAR can identify which subtests are worthy of inclusion in a reported score profile, while $\mathcal{G}$ can be used to determine the reliability of the score profile once the reportable subtests have been identified. As illustrated in the real data example presented earlier, VAR was employed to combine five subscores into three reportable subscores, and $\mathcal{G}$ was then obtained to document the reliability of the reconfigured score profile. In that example, narrowing from five to three subtests increased $\mathcal{G}$ for the reference group from .51 to .57, and increased $\mathcal{G}$ for the focal group from .71 to .83. Of course, this use of VAR and $\mathcal{G}$ is subject to the previously noted caveat that these indices refer to the quality of subscores for the population not for an individual examinee.

The results also suggest that $\mathcal{G}$ be routinely calculated just as a double check on VAR. A positive feature of $\mathcal{G}$ is that its theoretical basis—the ratio of true variance to observed variance—is relatively straightforward to understand and communicate to others. Both the present results and work by Feinberg and Jurich (2017) suggest that criterion of VAR $\geq$ 1 may be too liberal. Obtaining $\mathcal{G}$ alongside VAR may help temper the interpretation of results in those instances where VAR exceeds 1.0 but $\mathcal{G}$ is in the 50s or 60s. Alternatively, the present study also identified instances where VAR appeared to underestimate the utility of subscores due to the considerable variability in subscore means. In such instances, consideration might be given to relaxing the VAR criterion.

The present study had certain limitations which affect the generalizability of the results. As with any study in which item responses are simulated, the results are determined in part by the models used for generating item responses. To the extent that such models do not capture the nuanced sources of variance in the real-life data being simulated, the results would not generalize to practice. It was not believed to be a major limitation, as generally consistent results were obtained across two experiments and with a real data example. In addition, the present findings pertaining to VAR are in line with previous related studies (Sinharay, 2010). Another limitation was that using VAR as a basis for comparison required averaging it over subtests. While this interfered with the comparison of $\mathcal{G}$ and VAR in the pilot experiment where there were four and six subtests, the comparison of $\mathcal{G}$ with VAR was more useful in the second experiment where all conditions were limited to just two subtests. Nonetheless, it is still important to recognize that the two indices answer different questions.

One area of additional research would be to extend the work of Sinharay and Haberman (2014) by examining the invariance of $\mathcal{G}$ to score profiles for different demographic groups. Another would be to evaluate $\mathcal{G}$ for examinees at different levels of ability, as implied by the work of Haladyna and Kramer (2004) who reported that low scoring examinees exhibited more variable score profiles. Another follow-up is to estimate variance and covariance components of the $p^\bullet \times i^\circ$ using (a) the confirmatory factor analysis (CFA) framework suggested by Marcoulides (1996) and/or (b) the Bayesian framework suggested by Jiang and Skorupski (2017). Another line of research is to evaluate the reliability of score profiles aggregated at the level of the classroom or institution. Such an application of $\mathcal{G}$ would be a natural extension of Kane and Brennan's (1977) work on the generalizability of class means.

Appendix for Pilot Experiment: The Use of Multivariate Generalizability Theory to Evaluate the Quality of Subscores

Method

Design

This study evaluated the response of $\mathcal{G}$ to different conditions of subtest reliability, subtest correlations, and variation in subtest means. The fact that these factors are not independent (e.g., total test length determines the number of items per subtest and subtest reliabilities) prompted the authors to conduct three independent experiments where each study differed primarily in terms of the number of subtests and the reliability of those subtests. Within each study, conditions were created by completely crossing multiple levels of subtest reliability with three levels of subtest correlation and three levels of overall score profile variability. Total test reliability for all studies was high (low 90s). The three levels of population correlation, $\overline{\rho_{vv'}}$, were set at .70, .80, and .90. Three levels of population subtest means were created by varying the magnitude of the differences in subtest means, such that values of $var(\mu_v)$ were .06, .25, and .56. These values correspond to standard deviations in subtest means of 0.25, 0.50, and 0.75. Although this amount of variation in subtest means is large, it does occur in practice (e.g., Sinharay & Haberman, 2014).

While levels of $\overline{\rho_{vv'}}$ and $var(\mu_v)$ were identical across the three studies, the number of subtests and levels of subtest reliability,ρ_v^2, varied. Four levels of ρ_v^2 were created by manipulating the number of items per subtest for each study.

- Study A simulated a high reliability situation for which there is total test score partitioned into two correlated, quite reliable subtests (e.g., reading and mathematics; constructed response and selected response). The number of items per subtest was set at 100, 110, 120, and 130, resulting in four levels of ρ_v^2 equal to .85, .86, .87, and .88. Within a particular condition, both subtests had an equal number of items.
- Study B simulated moderately high levels of reliability for a test consisting of four subtests. Four levels of reliability were studied by creating subtests consisting of 60, 70, 80, and 90 items, producing levels of ρ_v^2 of .77, .79, .81, and .83. Within a particular condition, all four subtests had an equal number of items. The upper levels of reliability represent what might be encountered in a test battery developed with subscore interpretations in mind. The lower levels correspond to the types of subscores that might be seen on well-developed subtests for which subscores were intended to be useful but not for decision-making purposes. These values are toward middle and high end of the range of reliabilities investigated by Sinharay (2010).
- Study C reflected a situation where there are six subtests with each consisting of a modest number of items. It is not uncommon for tests in K-12 or credentialing to report several subscores, where subscores correspond to categories of the blueprint, and are reported simply because they are available and make conceptual sense. For Study C, the number of items per subtest was set at 35, 45, 55, and 65 producing conditions with ρ_v^2 equal to .66, .71, .75, and .78.

In summary, each of the three studies consisted of three levels of $\overline{\rho_{vv'}}$, three levels of $var(\mu_v)$, and four levels of ρ_v^2, for a total of 36 conditions per study. The levels of $\overline{\rho_{vv'}}$ and $var(\mu_v)$ were the same across studies, whereas the levels of ρ_v^2 were unique to that study. About 120 replications were run for each of the 36 × 3 = 108 conditions, with $N = 1{,}000$ simulated examinees per replication. The use of 100 replications is common practice (Feinberg & Wainer, 2014; Sinharay, 2010), and the present initial investigations found that 120 replications produced consistently small standard errors of the index of primary interest ($SE(\mathcal{G}) < .002$).

Item response simulation

Data were simulated to mimic the types of item responses obtained from a certification test, where subscore overreporting seems particularly prevalent (Puhan et al., 2010; Sinharay, 2010). Subscores were generated using a two-parameter, logistic multidimensional item response theory (MIRT) model (Haberman, von Davier, & Lee, 2008; Reckase, 2007). Let $\boldsymbol{\theta} = (\theta_1, \theta_2 \ldots \theta_k)$ correspond to the K-dimensional *true* ability parameter vector of an examinee. The probability of a correct response P to item i from an examinee can be expressed as

$$\frac{\exp(a_{1i}\theta_1 + a_{2i}\theta_2 + \cdots + a_{ki}\theta_k - b_i)}{1 + \exp(a_{1i}\theta_1 + a_{2i}\theta_2 + \cdots + a_{ki}\theta_k - b_i)},$$

where b_i is a scalar difficulty parameter, and $\mathbf{a}_i = (a_{1i}, a_{2i}, \ldots, a_{ki})$ is a vector of discrimination parameters of item i. As each item measures one subscore only, $\mathbf{a}_i$ can be specified as (0,…, a_{iV},…,0), where V is the identifier of the subscore. Each element in $\boldsymbol{\theta}$ can be regarded as a subtest in the current context, and θ_k is an examinee's score for subtest k. Item responses were generated by comparing P with a random draw u from a uniform distribution ranging from 0 to 1. If $P \geq u$, then the response x_i at item i is 1; otherwise if$p < u$, response $x_i = 0$.

Item discrimination and difficulty parameter estimates (a_i, b_i) were obtained from a physician certification test; these estimates were treated as known parameters and served as the basis

Table A1. Mean $\mathcal{G}$ and VAR Across Levels of Subtest Correlation ($\overline{\rho_{vv'}}$), Subtest Reliability (ρ_v^2), and *var*(μ_v).

			Mean $\mathcal{G}$				Proportion VAR > 1.0			
			var(μ_v)				*var*(μ_v)			
	$\overline{\rho_{vv'}}$	ρ_v^2	.06	.25	.56	*M*	.06	.25	.56	*M*
Study A	.70	.85	.67	.75	.82	0.75	1.00	1.00	1.00	1.00
		.86	.69	.77	.84	0.77	1.00	1.00	1.00	1.00
		.87	.71	.78	.85	0.78	1.00	1.00	1.00	1.00
		.88	.72	.80	.86	0.80	1.00	1.00	1.00	1.00
		M	0.70	0.78	0.84	0.77	1.00	1.00	1.00	1.00
	.80	.85	.59	.72	.80	0.70	.97	.97	1.00	0.98
		.86	.62	.73	.82	0.72	1.00	1.00	1.00	1.00
		.87	.64	.75	.83	0.74	1.00	1.00	1.00	1.00
		.88	.66	.76	.84	0.75	1.00	1.00	1.00	1.00
		M	0.63	0.74	0.82	0.73	0.99	0.99	1.00	0.99
	.90	.85	.48	.65	.77	0.64	.00	.00	.00	0.00
		.86	.51	.68	.80	0.66	.00	.00	.00	0.00
		.87	.53	.70	.81	0.68	.02	.01	.01	0.01
		.88	.55	.71	.82	0.69	.05	.04	.01	0.03
		M	0.52	0.68	0.80	0.67	0.02	0.01	0.01	0.01
Study B	0.70	.77	.55	.64	.74	0.65	.84	.87	.86	0.86
		.79	.60	.69	.77	0.69	.91	.91	.89	0.90
		.81	.63	.71	.79	0.71	.96	.91	.93	0.93
		.83	.65	.73	.81	0.73	.94	.92	.94	0.94
		M	0.61	0.69	0.78	0.69	0.91	0.90	0.91	0.91
	.80	.77	.48	.62	.72	0.61	.22	.28	.25	0.25
		.79	.52	.64	.75	0.64	.52	.48	.45	0.48
		.81	.54	.67	.77	0.66	.73	.71	.75	0.73
		.83	.57	.69	.79	0.68	.85	.90	.87	0.87
		M	0.53	0.66	0.76	0.65	0.58	0.59	0.58	0.58
	.90	.77	.38	.55	.69	0.54	.01	.00	.00	0.00
		.79	.41	.59	.72	0.57	.00	.00	.01	0.01
		.81	.44	.62	.74	0.60	.02	.02	.01	0.02
		.83	.46	.64	.77	0.62	.01	.01	.02	0.02
		M	0.42	0.60	0.73	0.58	0.01	0.01	0.01	0.01
Study C	.70	.66	.39	.47	.61	0.49	.35	.28	.30	0.31
		.71	.44	.55	.66	0.55	.47	.48	.48	0.48
		.75	.48	.59	.71	0.59	.50	.49	.49	0.50
		.78	.52	.63	.74	0.63	.50	.50	.50	0.50
		M	0.46	0.56	0.68	0.56	0.46	0.44	0.44	0.45
	.80	.66	.32	.43	.59	0.45	.01	.01	.00	0.01
		.71	.37	.5	.64	0.50	.04	.05	.06	0.05
		.75	.41	.54	.68	0.54	.22	.21	.20	0.21
		.78	.45	.58	.72	0.58	.41	.40	.39	0.40
		M	0.39	0.51	0.66	0.52	0.17	0.17	0.16	0.17
	.90	.66	.23	.38	.55	0.39	.00	.00	.00	0.00
		.71	.28	.43	.61	0.44	.00	.00	.00	0.00
		.75	.32	.48	.66	0.49	.00	.00	.00	0.00
		.78	.34	.53	.69	0.52	.00	.00	.00	0.00
		M	0.29	0.45	0.63	0.46	0.00	0.00	0.00	0.00

Note. VAR = value-added ratio

for simulating item responses. Certification test items are often easier and less discriminating than achievement and admissions test items, and that was the case here. The mean (and *SD*) of the discrimination parameters were 0.52 (0.24), whereas the corresponding values for the difficulty parameters were −1.49 (2.55).

True ability parameters of the examinees were assumed to follow a multivariate normal distribution whose mean vector is $\boldsymbol{\mu}$ and covariance matrix is $\boldsymbol{\Sigma}_p$. Both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_p$ were specified to meet the conditions shown in Table A1, where the number of elements in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_p$ is determined by the number of subtests. Specifically, four mean vectors of ability parameters $\boldsymbol{\mu}$ were specified to achieve the predetermined levels of between-subtest variance, $var(\mu_v)$, provided in Table A1. The mean of the elements in $\boldsymbol{\mu}$ was set to 0 for simplicity. As one example, for the six subtest condition where $var(\mu_v) = .25$, the elements of $\boldsymbol{\mu} = [0, .7, -.68, -.25, .45, -.22]$. The diagonal elements of $\boldsymbol{\Sigma}_p$ were constrained to be 1. The mean of the off-diagonal values are designated as $\overline{\rho_{vv'}}$ and correspond to the values in Table A1. However, the actual correlations for the true ability parameters were generated to be random variations of these target population values (i.e., the off-diagonal values were not constant from replication to replication).

Outcome variables

The two outcomes of interest are $\mathcal{G}$ and VAR. The suggestion of Feinberg and Wainer (2014) was followed, and VAR was computed from the two PRMSE values, such that if VAR > 1, then subscores add value and are worth reporting. Both $\mathcal{G}$ and VAR were obtained for each replication. Within each experimental condition, the mean $\mathcal{G}$ was reported across the 120 replications, as well as the proportion of replications for which VAR exceeds 1.0. Note that VAR is derived by comparing one subscore with the mean of subscores. Thus, for any test partitioned into v subtests, there will be v estimates of VAR. For purposes of this study, the average VAR was computed over all v subtests. The likely consequence is that any relationship between $\mathcal{G}$ and VAR could decline as the number of subtests increases.

Results

Table A1 provides detailed results for $\mathcal{G}$ and VAR for all three studies. For now, the authors of the present study focus their attention on Study B and Figure A1. Figure A1 depicts $\mathcal{G}$ as a function of subtest reliability for the three different levels of subtest correlation and variation in subtest means. Within each panel, it is evident that $\mathcal{G}$ increases with higher levels of subtest reliability and lower subtest correlations. Meanwhile, the plots across panels demonstrate the impact of profile variability on $\mathcal{G}$. For the left panel where score profiles are nearly flat, $\mathcal{G}$ does not exceed the mid-.60s. For the right panel where there are large differences in means, values of $\mathcal{G}$ range from about .70 to .80. Not only does $var(\mu_v)$ directly affect $\mathcal{G}$ as a main effect, but it also moderates the effect of subtest correlations, as evidenced by the closeness of the lines in the right panel. Specifically, large variation in subtest difficulty lessens the impact of subtest correlations on $\mathcal{G}$. To gauge the relative magnitude of the main and interaction effects, the variance in *cell means* represented in Figure A1 is partitioned. Most of the variation in means (95%) is explained by the three main effects, while a small portion (4%) can be attributed to the interaction between $var(\bar{\mu}_v)$ and $\overline{\rho_{vv'}}$.

The graphs for Studies A and C, which are not presented, followed the same pattern as the means presented in Figure A1, with the primary difference being either an overall upward shift in $\mathcal{G}$ for Study A or a downward shift for Study B. Table A1 provides detailed documentation of this overall shift across studies. Although not provided in the table, the grand means of $\mathcal{G}$ for Studies A, B, and C were 0.72, 0.64, and 0.51, respectively. The levels of $\mathcal{G}$ do not offer much

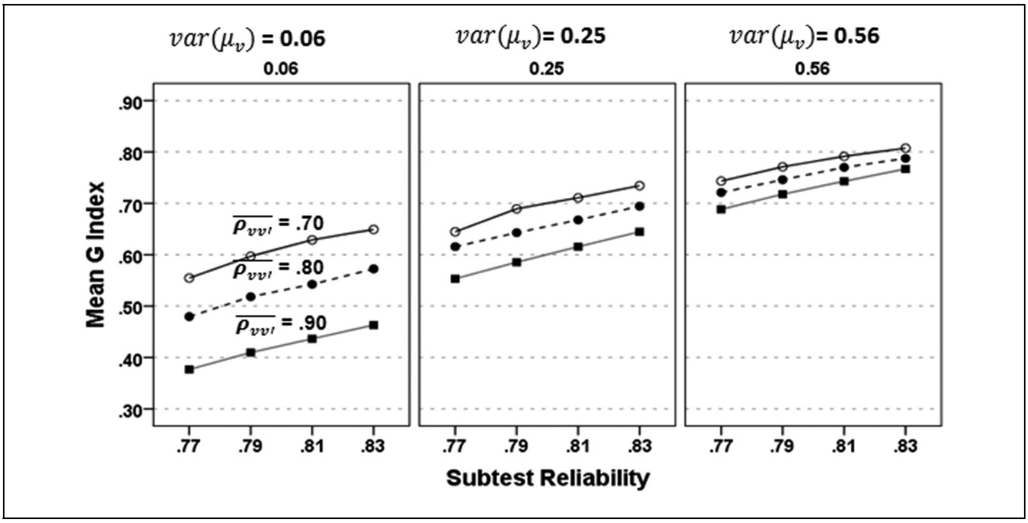

Figure A1. $\mathcal{G}$ as a function of subtest reliability (ρ_v^2) and mean correlation ($\overline{\rho_{vv'}}$) at three levels of variation in subtest means $var(\mu_v)$ for Study B.

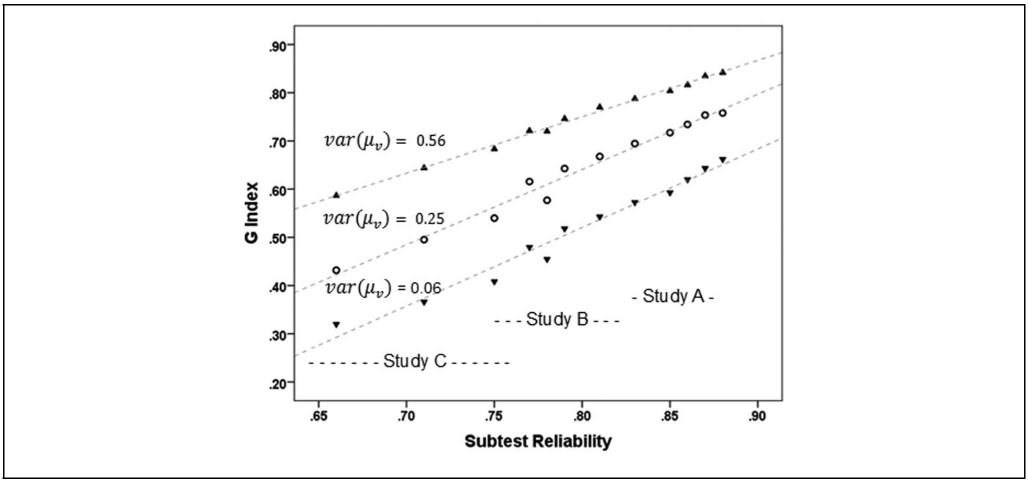

Figure A2. $\mathcal{G}$ as a function of subtest reliability and at three levels of variation for $\overline{\rho_{vv'}}$ = .80. Studies A, B, and C included two, four, and six subtests, respectively.

support for subscores at the modest levels of reliability of Study B, except when there is much variability in means. The results for Study A are more encouraging, particularly at modest levels of correlation ($\overline{\rho_{vv'}}$ = 70 or .80) and high levels of mean variation ($var(\bar{\mu}_v)$ = .25 or .56). By comparing column means for $\mathcal{G}$ across each of the three studies, it can be seen that variation in subtest means has the least effect on $\mathcal{G}$ for Study A, where the subtests are most reliable, and the greatest impact on $\mathcal{G}$ for Study C, where the subtests are least reliable.

Various plots that integrated results for $\mathcal{G}$ across studies were also examined. Figure A2 shows the results for all conditions at $\overline{\rho_{vv'}}$ = .80. The main point here is that results generalize across the three studies, at least within the levels of subtest reliability studied here. Figure A2 also suggests that results do not depend directly on the number of subtests studied but rather on the reliability of those subtests.

Table A1 provides results for VAR. When interpreting VAR for a single replication, the critical value is, of course, 1.0. However, when cumulating VAR across multiple replications within a condition, a critical value of .50 is suggested. That is, if the value in the table exceeds .50, then VAR is more likely than not to exceed 1.0 for those conditions. Study A resulted in VARs that would support the reporting of subscores under all conditions, where $\overline{\rho_{vv'}}$ was equal to .70 or .80, and for none of the conditions, where $\overline{\rho_{vv'}}$ = .90. For Study B, VAR approached 1.0 only when $\overline{\rho_{vv'}}$ = .70, and gave mixed results for the other conditions. For Study C, VAR indicated that subscores were worth reporting about half of the time when $\overline{\rho_{vv'}}$ = .70, but seldom reportable at $\overline{\rho_{vv'}}$ = .80 or .90.

It is apparent that VAR and $\mathcal{G}$ covary, although there are exceptions to this general trend. In Study A, for example, VAR appears to be generous, in that there are numerous conditions for which VAR is near 1.0, but the $\mathcal{G}$ indices are relatively low. For example, at $\overline{\rho_{vv'}}$ = .80, all 12 conditions have subscores deemed worth reporting based on VAR; however, the $\mathcal{G}$ indices for some of these conditions dip into the .50s and .60s. In contrast, VAR is not so liberal when subtests are highly correlated. Even though there are a few conditions at $\overline{\rho_{vv'}}$ = .90 where $\mathcal{G}$ exceeds .80, none of these conditions produced favorable VAR indices, confirming that VAR is not sensitive to variation in subscore means. This observation is also supported by the consistency in column means for VAR across levels of $var(\mu_v)$ k.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Zhehan Jiang https://orcid.org/0000-0002-1376-9439

References

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

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

Brennan, R. L. (2011). *Utility indices for decisions about subscores* (Research Report No. 33). Iowa, IA: Center for Advanced Studies in Measurement and Assessment.

Bridgeman, B. (2016). Can a two-question test be reliable and valid for predicting academic outcomes? *Educational Measurement: Issues and Practice*, *35*(4), 21-24.

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with college courses. *Journal of Educational Measurement*, *31*, 37-50.

Cronbach, L. J., & Gleser, G. (1953). Assessing similarity between profiles. *Psychological Bulletin*, *50*, 456-473.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, M. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Feinberg, R. A., & Jurich, D. P. (2017). Guidelines for interpreting 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.

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

Haberman, S. J., von Davier, M., & Lee, Y. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous distributions* (ETS Research Report No. RR-08-45). Princeton, NJ: Educational Testing Service.

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

Harris, D. J., & Hanson, B. A. (1991, April). *Methods of examining the usefulness of subscores*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostical assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60). Cambridge, UK: Cambridge University Press.

Jiang, Z., & Skorupski, W. (2017). A Bayesian approach to estimating variance components within a multivariate generalizability theory framework. *Behavior Research Methods*, 1-22. doi:10.3758/s13428-017-0986-3

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research*, *47*, 267-292.

Livingston, S. A. (2015). A note on subscores. *Educational Measurement: Issues and Practice*, *34*(2), 5.

Marcoulides, G. A. (1996). Estimating variance components in generalizability theory: The covariance structure analysis approach. *Structural Equation Modelling*, *3*, 290-299.

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education*, *23*, 266-285.

Reckase, M. D. (2007). Multidimensional item response theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 607-642). Amsterdam, The Netherlands: Elsevier Science B.V.

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

Sinharay, S. (2013). A note on assessing the added value of subscores. *Educational Measurement: Issues and Practice*, *324*(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*, 122-148.

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study for when the test is essentially unidimensional. *Applied Measurement in Education*, *23*, 63-86.

Thissen, D., Wainer, H., & Wang, X. B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement*, *31*, 113-123.

van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*, 339-356.