Investigating the Use of Multivariate Generalizability Theory for Evaluating Subscores

Zhehan Jiang

University of Kansas

Mark Raymond

National Board of Medical Examiners

Presented at the Annual Meeting of the
National Council on Measurement in Education April 2017

Author Note:

Zhehan Jiang, Doctoratal Candidate, School of Education, University of Kansas.

Mark Raymond, Research Director & Principal Assessment Scientist, National Board of Medical Examiners.

Abstract

Conventional methods for evaluating the utility of subscores rely exclusively on reliability and correlation coefficients. However, correlations can overlook an important source of variability: the variation in subtest means (i.e., subtest difficulty). As part of his treatment of multivariate generalizability theory, Brennan (2001) introduced a reliability-like index for score profiles 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. This simulation experiment investigated the properties of $\mathcal{G}$ under various conditions of subtest reliability, subtest correlations, and variability in subtest means. One main finding was that $\mathcal{G}$ indices typically were low; across all 224 conditions $\mathcal{G}$ ranged from .20 to .92, with an overall mean of .70. This finding is generally consistent with previous research indicating that subscores often do not have interpretive value. More specific findings indicate the conditions that produce favorable levels of$\mathcal{G}$. Importantly, there were many conditions for which conventional methods would indicate that subscores are worth reporting, but for which values of $\mathcal{G}$ fell into .60s and .70s. The results suggest that $\mathcal{G}$ can serve as a useful index for further characterizing the quality of subscores.

Keywords: subscore, multivariate generalizability theory, measurement, reliability, dimensionality, simulation

Investigating the Use of Multivariate Generalizability Theory for Evaluating Subscores

Examinees and other test users often expect to receive subscores in addition to total test scores (Huff & Goodman, 2007). This expectation is understandable: when students spend a few to several hours taking a test of different skill or content domains, it is only natural for those students, their parents, or teachers to wonder about their performance on specific skill or content domains. Indeed, the demand for subscores in K–12 settings has been fueled, in part, by government-mandated testing programs such as the No Child Left Behind Act of 2001 (NCLB) which call for diagnostic reports that allow legitimate stakeholders to use scores for remedial purposes (United States Department of Education, 2004). One limitation of subscores is that they often are not distinct from one another. A related problem is that the conceptual framework for assigning items to subtests may not hold up to empirical scrutiny; subject matter experts may have one framework in mind, but the data may suggest a different one (D'Agostino, Karpinski, & Welsh, 2011). Another problem is that subscores based on just a handful of test items often are not sufficiently reliable to allow one to generalize to the content domain that the items represent. The *Standards for Educational and Psychological Testing* (AERA et al., 2014) address the issue of subscores in multiple locations. For example, Standard 1.14 indicates that when interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretations should be provided (p. 27). The commentary for that Standard goes on to note that the testing agency is responsible for demonstrating the distinctiveness and reliability of subscores.

One straightforward approach for evaluating the distinctiveness of subscores is to compute correlations among subscores or between subscores and the total score (e.g., Haladyna & Kramer, 2004). A correlation near 1.0 between any two subscores implies that the two scores measure a single dimension, trait, or proficiency. Just how low the correlation should be to support the reporting subscores is a matter of judgment. Factor analytic methods also have been used to investigate the utility of subscores (Haladyna & Kramer, 2004; Stone, Ye, Zhu & Lane, 2010). One common strategy is to inspect the eigenvalues of the correlation matrix and assume that the number of distinct traits corresponds to the number of eigenvalues greater than 1.0 (e.g., Sinharay, Haberman, & Puhan, 2007). Alternatively, one can examine the relative magnitudes of eigenvalues and assume a single total score is sufficient if the first eigenvalue overwhelms all others (Stone et al., 2010). Confirmatory factor analysis lends itself to the use of more formal criteria by specifying alternative models (e.g., based on three, two, or a single score), and then inspecting factor loadings and fit indices to determine which subscore model provides the best fit (e.g., D'Agostino et al., 2011; Thissen, Wainer, &Wang, 1994). Although factor analysis provides a structured approach to analyzing correlations, deciding whether to report subscores still is a matter of judgment or of applying somewhat arbitrary criteria.

A method developed by Haberman (2008) removes the subjectivity of interpreting correlations, factor loadings, and related indices. Furthermore, his method incorporates both subscore distinctiveness and subscore reliability into a single decision rule about whether to report subscores. The method is based on the principle 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-squared 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. They further found through simulation studies that *VAR* could be reasonably well estimated from a simple linear model that includes the reliability coefficients of the subscore and total scores, and the disattenuated correlation between the subscores and total test scores (Feinberg & Wainer, 2014). However, both VAR and PRMSE as originally proposed require that correlations between subscores and total scores be computed based on the remainder score described above, which can be a bit tedious. Brennan (2011) proposed a utility index, *U*, which produces the same decisions as PRMSE and VAR but does *not* rely on regressed estimates of true scores or on remainder score correlations. Utility is easily estimated from observed variances, covariances and reliability coefficients. However, Brennan's (2011) formulation is presented only in an unpublished technical report and has received little attention.

The overwhelming 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 over-reporting of subscores seems particularly prevalent on licensure and certification tests. For example, a study of six licensure tests found that only one of a total of 24 subscores satisfied the PRMSE criterion (Puhan et al., 2010). Another study of a wide variety of testing programs found that subscores were useful for 16 of 92 subscores reported across 24 different tests (Sinharay, 2010).

Although correlational methods are a useful way to summarize relationships among variables, they have certain limitations. First, correlations essentially eliminate differences in means and variances across variables, causing one to overlook potentially important systematic differences in subtest difficulty (Cronbach & Gleser, 1953). Second, conventional methods for evaluating subscore utility address whether subscores are reliably different from the total score. However, test users are often interested in questions like, “Are these two subscores different from each other?” Or, “To what extent does my score profile reliably indicate my strength and weaknesses?” A third limitation is that correlations are relatively insensitive to substantial changes in score profiles subgroups of examinees (Livingston, 2015). Consider two subscores, X and Y, where r_{xy}= .750. Adding one-half of a standard deviation to 50% of the scores on Y has minimal impact on the correlation, changing it to r_{xy}= 0.726. To make this example more concrete, consider the general finding that females score higher than males on essay questions and lower on multiple-choice questions; while profiles means based on gender clearly differ, the correlations do not, thereby suggesting that essay scores should not be separately reported, although not doing so overlook important differences (Bridgeman & Lewis, 1994).

There is growing recognition that the analysis of between-subject covariation (i.e., correlations) is not sufficient for understanding test dimensionality, and that it is necessary to also study within-subject variation (van der Maas, Molenaar, Maris, Kievit & Borsboom , 2011). Because score profiles are what testing agencies report and what test users interpret, it only seems natural to inspect the properties of actual score profiles when evaluating the utility of subscores. The guiding principle is that score profiles must contain variability in order to be informative. Flat score profiles contain no information above that provided by the total score, while variable profiles may contain additional information. The challenge is to differentiate signal from noise in the score profile.

Multivariate Generalizability Theory and $\mathcal{G}$-index

Cronbach et al (1972) laid the groundwork for differentiating signal from noise in score profiles within the context of multivariate generalizability theory (G-theory). Their efforts remained only 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, p. 323). One important difference between $\mathcal{G}$ and correlational methods is that it treats the profile as the unit of analysis. That is, $\mathcal{G}$ is characterizes entire score profiles rather than specific subtests.

The G-theory design most relevant to the study of subscores involves a different set of items (i) being assigned to each of several subtest (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 $p^{\bullet} \times i^{\circ}$, 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., with each subtest), while the open circle indicates that items are not linked across the different subtests (i.e., each subtest consists of a unique set of items).

A multivariate G study based on the $p^{\bullet} \times i^{\circ}$ design produces matrices of variance-covariance components for persons, items, and error, designated as $\boldsymbol{\Sigma}_p$, $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Sigma}_\delta$. Brennan (2001) defines the generalizability index for score profiles as:

$$\mathcal{G} = \frac{\mathcal{V}(\mu_p)}{\mathcal{V}(\bar{X}_p)} \tag{1}$$

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. Brennan (2001) provides an estimate of the numerator as:

$$\mathcal{V}(\mu_p) = [\overline{\sigma_v^2}(p) - \overline{\sigma_{vv\prime}}(p)] + var(\mu_v)\,, \tag{2}$$

where

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

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

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

The denominator for equation (1) is obtained by:

$$\mathcal{V}(\overline{X}_p) = [\overline{S_v^2}(p) - \overline{S_{vv\prime}}(p)] + var(\overline{X}_v) \tag{3}$$

where

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

$\overline{S_{vv\prime}}(p)$ = mean of the all n_v elements in **S** .

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

Note that $var(\mu_v)$ is estimated by $var\left(\overline{X}_v\right)$. One computational convenience is that **S** can be obtained either by summing variance-covariance component matrices $\mathbf{\Sigma}_p$ and $\mathbf{\Sigma}_\delta$ or by computing it directly from the observed scores. Another convenience is that for the $p^\bullet \times i^\circ$ design, the covariance components for observed scores provides an unbiased estimate of covariance components for universe score. That is, $\sigma_{vv'}$ = $S_{vv'}$, which means that the off-diagonal elements of $\boldsymbol{\Sigma_p}$ = **S**.

The numerator and denominator of equation (1) have a few noteworthy features. First, the principle difference between them is that equation (2) contains universe score variances and covariances, while equation (3) contains observed variances and covariances. Second, covariances are subtracted out of both the numerator and denominator. Thus, as subscore correlations approach 1, the difference between $\overline{\sigma_v^2}(p)$ and $\overline{\sigma_{vv'}}(p)$ approaches 0, as does the difference between $\overline{S_v^2}(p)$ and $\overline{S_{vv'}}(p)$; in both instances an increase in subtest correlations decreases $\mathcal{G}$. Third, it is evident that differences in subtest difficulty contribute to $\mathcal{G}$. If subscores all have the same mean, then $var(\overline{X}_v)$ = 0 and subtest difficulty will have no effect. Fourth, if subtests correlate zero on average and all subtest means are equal, equations (2) and (3) reduce to $\overline{\sigma_v^2}(p)$ / $\overline{S_v^2}(p)$, which is the ratio of true score variance to observed score variance, and is the conventional definition of reliability. Under these conditions $\mathcal{G}$ equals the average of the subtest reliabilities. In other words, subtest reliability places an upper limit $\mathcal{G}$ when means are equal, and it can only diminish as correlations increase. A fifth and related corollary is that $\mathcal{G}$ can exceed subtest reliability only as subtest correlations approach zero and the variance in subtest means is greater than zero.

To date, $\mathcal{G}$ has received little attention in the literature and there is no practical guidance regarding its interpretation. Brennan (2001) provides an example where $\mathcal{G}$ is computed for three subtests that make up 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 subtest 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. The purpose of this study is to provide an initial investigation into the properties of $\mathcal{G}$. More specifically, we conduct a series of experiments using simulated test score data to evaluate the sensitivity of $\mathcal{G}$ to different test conditions of theoretical interest and likely to be encountered in practice. To provide a context for interpreting $\mathcal{G}$, we compare it to PRMSE.

Methods

Study Designs

The simulation was designed to evaluate the response of $\mathcal{G}$ to different conditions of subtest reliability, total test reliability, subtest correlations, and variation in subtest means. The first three of these factors have been studied in prior simulations of subscore utility, while the last has not. It is evident that the factors of interest are not independent. For example, for a total test of a specified length, the number of item per subtest and subtest reliabilities will be determined in part by the number of subtests. Given the lack of independence, we conducted three independent experiments where each study differed primarily in terms of the number of subtests and the reliability of those subtests. Within each experiment, conditions were created by completely crossing multiple levels of subtest reliability with four levels of subtest correlation and four levels of overall score profile variability. Subtest reliability, ρ_v^2, was controlled by manipulating the number of items per subtest as specified in Table 1. The four levels of population correlation, $\overline{\rho_{vv'}}$, were set at .60, .70, .80, and .90. Meanwhile the four 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 0.06, 0.25, 0.56, and 1.00. Although levels of subtest variability greater than 0.25

likely exceed what is commonly encountered in practice, given that this is an initial study of $\mathcal{G}$, it will be useful to document how it performs at extreme conditions of $var(\mu_v)$, particularly since that is what most distinguishes $\mathcal{G}$ from other methods.

[INSERT TABLE 1 ABOUT HERE]

Table 1 summarizes the study design and levels for each factor. Study A consisted of two subtests and corresponds to situations for which there is total test score partitioned into two correlated, quite reliable subtests (e.g., reading and mathematics; constructed response and selected response) with the values of ρ_v^2 ranging from .85 to .89. Within a particular condition, both subtests had an equal number of items. Study B included four subtests with five levels of ρ_v^2 that ranged from .73 to .83. The upper levels of reliability for study B 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. For example, even reliability coefficients of .73 are above the median subtest reliability reported by Sinharay (2010). Study C has six subtests. 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. Study C examines conditions where a total test is split into six subtests consisting of with reliability coefficients ρ_v^2 ranging from .65 to .78. Studies A and B each consisted of 80 conditions, while Study C consisted of 64 conditions, for a total of 224 experimental conditions across the three studies. We ran 120 replications for each condition with N = 1000 simulated examinees per replication.

Item Response Simulation

Subscores were generated using a two-parameter, logistic multidimensional item response theory (MIRT) model (Reckase, 2007; Haberman, von Davier, & Lee, 2008). 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 $\boldsymbol{a_i} = (a_{1i}, a_{2i}, \ldots, a_{ki})$ is a vector of discrimination parameters of the item *i*. 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 for simulating item responses. Certification tests often are found to be easier and less discriminating than achievement and admissions tests (Raymond & Luecht, 2013), and that was the case here. The mean (and SD) of the discrimination parameters were .52 (.24), while 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 in Table 1, 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, $var(\overline{\mu}_v)$, provided in Table 1. The mean of the elements in $\boldsymbol{\mu}$ was set to 0 for simplicity. As one example, for the two condition case where $var(\mu_v)$ = .06, the elements of $\boldsymbol{\mu}$ = [0.178,-0.178]. As another example, for the six subtest condition where $var(\mu_v)$ = .25, the elements of $\boldsymbol{\mu}$ = [0,0.7,-0.68,-0.25,0.45,-0.22]. The diagonal elements of $\boldsymbol{\Sigma}_p$ were constrained to be 1 (i.e., correlation matrix). The mean of the off-diagonal values are designated as $\overline{\rho_{vv\prime}}$ and correspond to the values in Table 1. However, the actual correlations for the true ability parameters were generated to be random variations of these target population values. That is, for each specified level of average correlation, the off-diagonal values were not constant from replication to replication. For example, in the Study C condition where $\overline{\rho_{vv\prime}}$ = 0.6, the six actual correlations used for response generation could take on the values such as $\boldsymbol{\Sigma}_p$ = [.40, .80, .60, .50, .60, .70], or values something like [.35, 75, 45, 85, .65, .55]. Creating random variation in correlations ensured that the average level of correlation was equal to $\overline{\rho_{vv\prime}}$, while allowing actual subtest correlations to be different without constraining the elements $\boldsymbol{\Sigma}_p$ to a specific pattern of correlations.

Outcome Variables

The two outcomes of interest are $\mathcal{G}$ and proportion reduction in mean-squared error (PRMSE) (Haberman, 2008). $\mathcal{G}$ was computed according to equations (1) through (3). PRMSE for total test scores and subtest scores was computed according to the method described by Haberman (2008). In addition, we followed the suggestion of Feinberg and Wainer (2014), and computed a value added ratio (VAR) from the two PRMSE values such that if VAR > 1, then subscores add value and are worth reporting for that replication. Both $\mathcal{G}$ and VAR were computed for each replication. For all studies, within each experimental condition we report the mean $\mathcal{G}$ across the 120 replications, as well as the proportion of replications for which VAR exceeds 1.0. The intent is not to compare $\mathcal{G}$ and VAR, but to use our familiarity with VAR to help interpret $\mathcal{G}$.

Result

We first present high level results across all three studies, next describe the detailed findings for one study, and then assemble findings from all three into a whole that can be generalized to other contexts. Our intent is to give the reader a thorough understanding of the results without presenting detailed outcomes for each study.

[INSERT TABLE 2 ABOUT HERE]

Table 2 summarizes both $\mathcal{G}$ and VAR for all 3 studies for different conditions of $var(\overline{X}_v)$ and $\overline{\rho_{vv'}}$ while collapsing across average ρ_v^2. Subtest reliability ρ_v^2 was chosen to collapse because it contributed least to main and interaction effects on $\mathcal{G}$. It can be seen in left portion of Table 2 that within each study $\mathcal{G}$ increases as the subtest correlations decrease and variation in subtest means increases. Across studies $\mathcal{G}$ declines with decreases in ρ_v^2 as evidenced by lower values of $\mathcal{G}$ for studies B and C. A notable finding is that $\mathcal{G}$ tends to be low, even for conditions where one might expect it to be high, such as in the conditions of Study A where the subtest reliabilities are moderately high.

The right section of Table 2 summarizes the proportion of replications for which VAR > 1.0. When interpreting VAR for a single replication, the critical value is, of course, 1.0. However, when cumulating VAR indices across multiple replications, we suggest a critical value of .50. 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 most conditions studied. The only study A conditions for which VAR was consistently low was for subtest correlations of .90. In contrast, none of the conditions in Study B or Study C consistently met the VAR criterion for reporting subscores. This is a consequence of the lower levels of ρ_v^2 for studies B and C. Table 2 also indicates that VAR changes more dramatically than

$\mathcal{G}$ (i.e., larger decreases) as subtest correlations decrease and as one moves down the table from Study A to Studies B and C.

[INSERT FIGURE 1 ABOUT HERE]

To show the effect of subtest reliability on $\mathcal{G}$, and to further illustrate the relationship between $\mathcal{G}$ and VAR, we provide detailed results for Study A. Figure 1 depicts mean $\mathcal{G}$ as a function of subtest reliability for the different levels of subtest $\overline{\rho_{vv\prime}}$ and $var(\overline{\mu}_v)$. Within each panel it is evident that $\mathcal{G}$ is responsive to both subtest reliability and the correlations among subtests, while results across panels demonstrate the impact of profile variability on $\mathcal{G}$. In panel A of Figure 1, where there is little difference subtests means, $\mathcal{G}$ ranges from .48 to .79, with an overall average of .66. Meanwhile, for panel D, where there are very large differences in means, values of $\mathcal{G}$ average about .89. Not only does $var(\overline{\mu}_v)$ impact $\mathcal{G}$, but it also moderates the impact of subtest correlations and subtest reliability on $\mathcal{G}$. This is evidenced by the nearly overlapping lines in panel D and the reduced slopes for those lines compared to the steeper slopes and more variable intercepts in panels A, B, and C.

[INSERT FIGURE 2 ABOUT HERE]

Figure 2 summarizes the proportion of replications for which VAR > 1.0. Again we suggest that if the plotted value exceeds .50 for a given condition, then VAR indicates that subscores generally are worth reporting for that condition. Panels A through D in Figure 2 correspond to the same four conditions presented in Figure 1. The four panels in Figure 2 have a nearly identical appearance because $var(\overline{\mu}_v)$ has no direct impact on VAR (although there minor indirect effects associated with changes in reliability and covariance). Also note that VAR suggests that subscores are worth reporting for all conditions except when subtest correlations are .90. Even the conditions for which $\mathcal{G}$ was in the mid .50s and.60s (Figure 1A), VAR exceeded 1.0 most of time (Figure 2A).

Graphs for Studies B and C are not presented because they generally parallel those for Study A. There is, however, one important distinction: the values of $\mathcal{G}$ and VAR in Studies B and C are consistently lower than for Study A, which is seen in the lower lines (i.e., smaller intercepts) for studies B and C. This is a consequence of the shorter and less reliable subtests for studies B and C – a result that also is evident in Table 2. The slopes of the lines were very similar across studies A, B, and C suggesting that interactions of ρ_v^2 with $var(\overline{\mu}_v)$ and $\overline{\rho_{vv\prime}}$ were comparable for the three studies.

[INSERT FIGURE 3 ABOUT HERE]

The results presented so far suggest that VAR and $\mathcal{G}$ covary. We investigated this relationship by producing scatterplots for $\mathcal{G}$ and VAR within each of the three studies at each of the four levels of $var(\overline{\mu}_v)$. There were 12 such scatterplots. Figure 3 presents the plots for Study A. A single marker in Figure 3 corresponds to the values of $\mathcal{G}$ and VAR for each of 20 experimental conditions (five levels of reliability x four levels of correlation). It is evident that the two indices convey similar types of information about subscore utility. However, the relationship between $\mathcal{G}$ and VAR deteriorates as $var(\overline{X}_v)$ increases because $\mathcal{G}$ becomes consistently high. Across the four panels, VAR consistently identifies that subscores be reported for about 15 of the 20 conditions. However, that number would vary for $\mathcal{G}$ across the four panels, depending on the threshold value that one establishes for $\mathcal{G}$. If, for example, one arbitrarily established a threshold of .75 for $\mathcal{G}$, subscores would be deemed worth reporting for only 3 of the 20 conditions in Figure 1A. Those values for Figures 1B through 1D would be 13 of 20 conditions, 20 of 20 conditions, and 20 of 20 conditions, respectively. If one narrows the comparison to those which are more likely to be seen in practice, levels of $\mathcal{G}$ are not high. In Figures 1A and 1B, for example, VAR consistently indicated that subscores had utility for all conditions except those where subtest correlations were.90. Meanwhile, G indices for these same 30 conditions ($var(\overline{\mu}_v)$ = .06, .25; ρ_v^2 = .85, .86, .87, .88, .89; and $\overline{\rho_{vv\prime}}$ = .60, .70, .80), $\mathcal{G}$ ranged from .59 to .84, with an average of .74. The results suggest that there are instances where VAR may be satisfactory but for which $\mathcal{G}$ might regarded as low enough to regard subscores with suspicion.

We also computed Pearson correlation coefficients between $\mathcal{G}$ and VAR based on the plots in Figure 3 (i.e., for Study A). The correlations were .84, .73, .60, and .44 across the levels of $var(\overline{\mu}_v)$. For study B, the correlations ranged from .92 to .55 across the four levels of $var(\overline{\mu}_v)$, while for study C, they decreased from .90 to .50. These trends make sense: $\mathcal{G}$ and VAR are expected to diverge as subtest means become more variable given that the primary difference between the two indices is that VAR does not consider differences in subtest means.

[INSERT FIGURE 4 ABOUT HERE]

Figure 4 integrates results for $\mathcal{G}$ across the three studies. To facilitate interpretation and reduce clutter, the least realistic conditions are excluded from the figure, that is, those conditions where $\overline{\rho_{vv\prime}}$ =.60. This exclusion is consistent with previous reviews indicating that disattenuated subtest correlations are more typically in the .70s, .80s and .90s (Sinharay, 2010; Sinharay & Haberman, 2014). Figure 4 confirms the results presented in Table 2 and Figure 1, namely that $\mathcal{G}$ increases with increasing levels of subtest reliability and greater differences in subtest means. The figure also confirms that variation in subtest means also moderates the influence of subtest reliability and correlations on $\mathcal{G}$. More specifically, large variation in subtest difficulty lessens the impact of subtest reliability and correlations on $\mathcal{G}$. Importantly, Figure 4 also suggests that results generalize across studies, at least within the levels of subtest reliability studied here, and also suggest that results do not depend directly on the number of subtests studied, but rather on the empirical properties of those subtests.

Discussion

The simulations produced several findings that shed light on the potential utility of $\mathcal{G}$ for applied work. As expected, $\mathcal{G}$ increased with higher levels of subtest reliability, greater differences in subtest means, and lower levels of subtest correlation. In addition, greater variation in subtest means lessened the impact of subtest correlations and subtest reliability on $\mathcal{G}$. As suggested by previous research (Sinharay, 2010), the number of subtests did not seem to matter much. That is, for a given level of subtest reliability and correlation, the findings observed for the two subtest conditions generalized to the findings for the four and six subtest conditions. The number of subtests had only an indirect impact on $\mathcal{G}$ in that increasing the number of subtests lessens the number of items available for each subtest, resulting in less reliable subscores.

One of the more interesting findings was that $\mathcal{G}$ seldom reached conventionally-acceptable levels of reliability. Only under the most favorable conditions did $\mathcal{G}$ approach or exceed .80; while under more common conditions it fell into the .40s, .50s and .60s. For example, it was shown that where levels of $var(\overline{\mu}_v)$ are only modest (Figure 1, panels A and B), $\mathcal{G}$ exceeded .80 only when subtest reliability reached the mid .80s and subtest correlations were at .60 and .70. These conditions are seldom met by the types of subscores often reported in practice; instead, it is more typical to see subtest reliabilities in the .70s and correlations in the .80s and .90s (e.g., Sinharay, 2010; Sinharay & Haberman, 2014). It appears that the low values of $\mathcal{G}$ observed here are due in large part to its sensitivity to subtest reliability. When there is no variability in subtest means, subtest reliability imposes an upper limit on $\mathcal{G}$, and the index can only decline from that value as subtest correlations exceed zero, as they almost always do.

It is common to see little or no variability in subtest means. In fact, many testing programs scale subtests to have equal means. However, in those instances where score profiles deviate from flatness, $\mathcal{G}$ can make a contribution to understanding the utility of score profiles. For example, in a study of the utility of subscores for demographically distinct groups of examinees (e.g., gender, language), the score profiles presented by Sinharay and Haberman (2014) demonstrated considerable variability for some of the tests, even though the total group score profiles were flat. In such instances, the variation in means for some subgroups would contribute $\mathcal{G}$ and may indicate differential reliability of score profiles for subgroups. The ability of $\mathcal{G}$ to detect group differences could prove to be a useful area for further inquiry given that correlations alone may not be sensitive to such differences (e.g., Bridgeman & Lewis, 1994; Livingston, 2015).

The low $\mathcal{G}$ indices observed here are at once disconcerting and yet consistent with expectations. On one hand, given that the recent literature on PRMSE (VAR) has shown that subscores usually are not very informative, one should expect to see low values of $\mathcal{G}$. On the other hand, it was surprising that there were numerous conditions for which VAR indicated that subscores should be reported but $\mathcal{G}$ indices were cautiously low. Across all conditions where VAR indicated that subscores added value, the mean $\mathcal{G}$ was .76. More realistic conditions (e.g., Figures 1A and 1B) indicated even lower values of $\mathcal{G}$. It was shown that if one established an arbitrary value of $\mathcal{G} \geq .75$ for declaring that subscore profiles are sufficiently reliable to report, then only 40% of the conditions in Figures 1A and 1B would have met that threshold. While $\mathcal{G}$ and VAR are strongly related, they clearly provide different types of information. Just knowing that subscores are worth reporting according to VAR may not be sufficient; it also is useful to know what percent of the variability in reported score profiles can be regarded as reliable variance. As the present results indicate, using only VAR may present an overly optimistic view of subscores.

To be useful in practice, $\mathcal{G}$ requires guidelines for interpretation, and such guidelines will evolve only as users gain experience with real and simulated data. One can imagine two very different approaches to using $\mathcal{G}$. One approach is to propose generally accepted thresholds for $\mathcal{G}$ below which subscores would not be reported. This might be done by relying on VAR for guidance. For example, across simulation studies like this, one might identify the values of $\mathcal{G}$ that would maximize the agreement between it and VAR, and propose such a value as the minimum $\mathcal{G}$ index required for subscore reporting. Or, one could establish threshold on more theoretical grounds. For example, one might adopt the position that for low-stakes uses of subscores, the score profile should have twice as much signal as noise. This would correspond to a $\mathcal{G}$ of .67. One could imagine using similar logic to set higher thresholds (e.g., .80) for higher stakes decisions. A second and perhaps more productive way to use $\mathcal{G}$ is not to seek a threshold, but to use it in conjunction with VAR. With this approach, VAR would be used to decide whether subscores should be reported, while $\mathcal{G}$ would then be provided to summarize the reliability of the reported score to profile. One could imagine situations where $\mathcal{G}$ might moderate the decision. For example, if VAR was near 1.0 but $\mathcal{G}$ was sufficiently high due to considerable variability in subscore means, then subscores might be reported anyway. All of these uses must recognize that $\mathcal{G}$ is a single index for characterizing the quality of an entire score profile, while VAR is used for gauging the quality of and making decisions about each subtest within a score profile.

This first study on the use of $\mathcal{G}$ for characterizing the quality of score profiles suggests areas for additional research. One would be to examine the invariance of $\mathcal{G}$ to score profiles for different demographic groups of examinees or for examinees at different levels of ability. Such studies are suggested Sinharay and Haberman (2014) who identified a few instances for which subscores were found to be worth reporting for some groups of examinees but not for others, and by Haladyna and Kramer (2004) who observed more variable score profiles for low-scoring examinees. Another line of research might 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 within a univariate framework. Finally, since this study evaluated simulated data only for a limited number of conditions, additional work expand the range of conditions investigated, or study finer distinctions between levels of a given factor.

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.

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

D'Agostino, J., Karpinski, A., & Welsh, M. (2011). A method to examine content domain structures. *International Journal of Testing*, *11*, 295-307.

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. & 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*(2), 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*(4), 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.

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

Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge, MA: Harvard University Press.

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

Puhan, G., Sinharay, S., Haberman, S., & Larkin, K. (2008). *Comparison of subscores based on classical test theory methods*. ETS Research Report No. RR-08-54). Princeton, NJ: Educational Testing Service.

Raymond, M. R., & Luecht, R. M. (2013). Licensure and certification testing. In K. Fl.Geisinger, B. A. Bracken, J. F. Carlson, J. I. C. Hansen, N. R.Kuncel, S. P. Reise, & M.C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology: Vol. 3.Testing and assessment in school psychology and education* (pp. 391–414). Washington, DC: American Psychological Association

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: North Holland

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.

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*, 63086.

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(2)*, 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(2)*, 339–356.

Yao, L. H., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31*(2), 83-105.

United States Department of Education (2004). Testing for results: Helping families, schools, and communities improve student achievement. In *NCLB / Stronger Accountability (Introduction)*. Retrieved June 12, 2016 from http://www2.ed.gov/nclb/accountability/ayp/testingforresults.html

Table 1. Summary of Experimental Design

Study	**No. of subtests**	**subtest length and reliability,** ρ_v^2	**Variance in Subtest Means,** $\boldsymbol{var(\bar{\mu}_v)}$	**Mean Subtest Correlation,** $\overline{\boldsymbol{\rho_{vv'}}}$
A	2	100, 110, 120, 130, 140 .85, .86, .87, .88, .89	0.06, 0.25, 0.56, 1.00	.60, .70, .80, .90
B	4	50, 60, 70, 80, 90 .73, .77, .79, 81, .83	0.06, 0.25, 0.56, 1.00	.60, .70, .80, .90
C	6	35, 45, 55, 65 .66, .71, .75, .78	0.06, 0.25, 0.56, 1.00	.60, .70, .80, .90

Table 2. Mean Values of $\mathcal{G}$ and VAR across levels of $\overline{\boldsymbol{\rho}_{\boldsymbol{vv'}}}$ and $var(\overline{\mu}_v)$ within each study while collapsing across levels of subtest reliability.

		Mean $\mathcal{G}$ Index				**Proportion VAR > 1.0**			
	$\overline{\boldsymbol{\rho}_{\boldsymbol{vv'}}}$	variance in subtest means, $var(\overline{\mu}_v)$				variance in subtest means, $var(\overline{\mu}_v)$			
		0.06	0.25	0.56	1.00	0.06	0.25	0.56	1.00
Study	.60	0.756	0.814	0.862	0.902	1.00	1.00	1.00	1.00
A*	.70	0.709	0.783	0.849	0.895	1.00	1.00	1.00	1.00
	.80	0.638	0.749	0.831	0.887	0.99	0.99	1.00	1.00
	.90	0.528	0.694	0.807	0.877	0.04	0.03	0.03	0.05
	mean	0.658	0.760	0.837	0.890	0.76	0.76	0.76	0.76
Study	.60	0.644	0.715	0.783	0.837	0.94	0.94	0.94	0.93
B*	.70	0.588	0.678	0.763	0.829	0.89	0.88	0.88	0.88
	.80	0.510	0.639	0.740	0.816	0.48	0.49	0.48	0.47
	.90	0.403	0.583	0.712	0.804	0.01	0.01	0.01	0.01
	mean	0.536	0.654	0.750	0.822	0.58	0.58	0.58	0.57
Study	.60	0.514	0.596	0.700	0.775	0.50	0.49	0.50	0.50
C*	.70	0.457	0.558	0.678	0.763	0.46	0.44	0.44	0.44
	.80	0.387	0.511	0.658	0.755	0.17	0.17	0.16	0.15
	.90	0.294	0.453	0.627	0.738	0.00	0.00	0.00	0.00
	mean	0.413	0.530	0.666	0.758	0.28	0.28	0.28	0.27

* The mean subtest reliability coefficients for Studies A, B, and C were .0.87, 0.79, and 0.72, respectively.

Figure 1. Mean $\mathcal{G}$ as a function of subtest reliability for the different levels of subtest $\overline{\boldsymbol{\rho}_{vv'}}$ and $var(\overline{\mu}_v)$ for Study A (2 subtests).

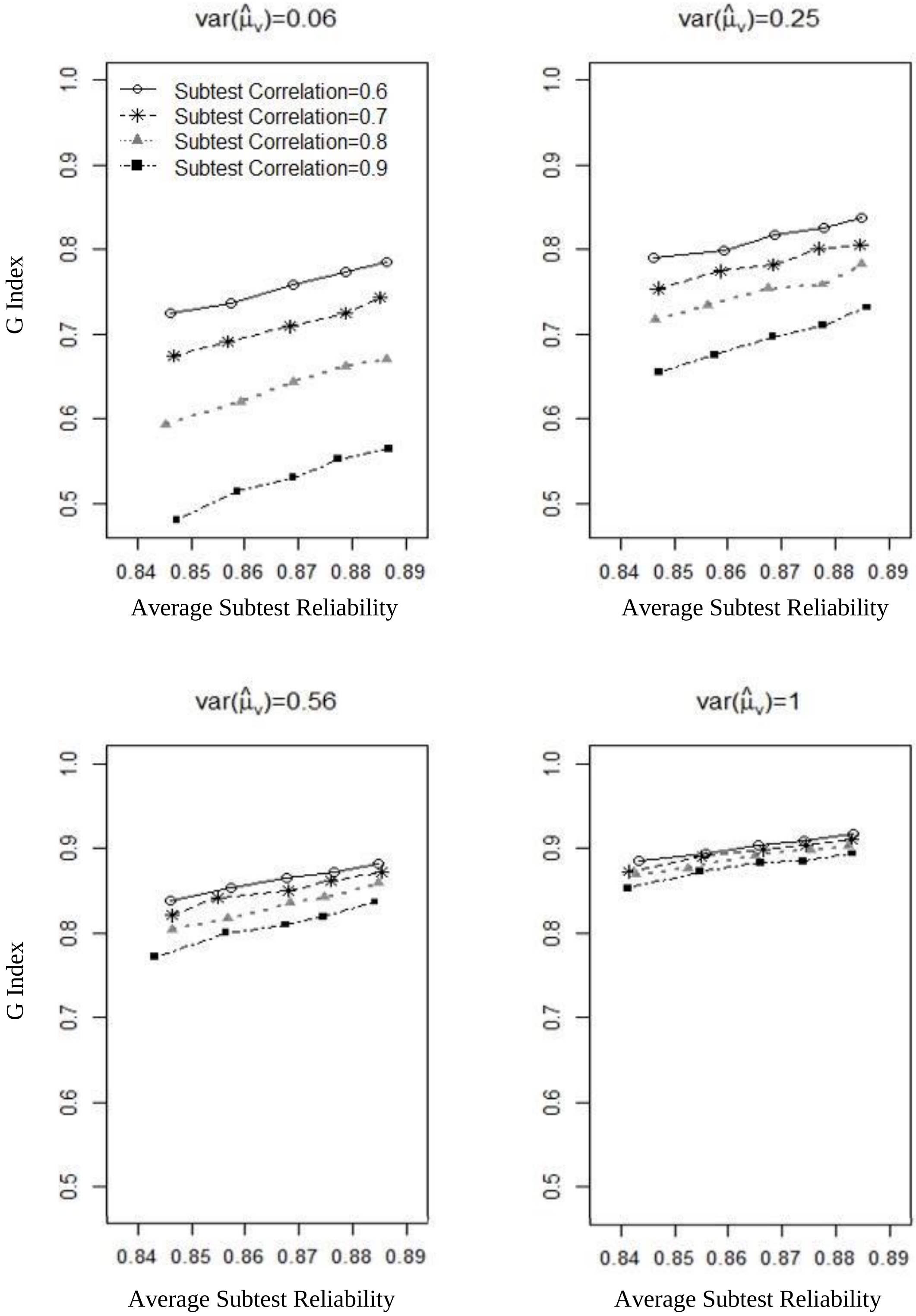


Figure 2. Proportion of replications where VAR $\geq$ 1.0 as a function of subtest reliability for the different levels of subtest $\overline{\boldsymbol{\rho}_{vv'}}$ and $var(\overline{\mu}_v)$ for Study A (2 subtests).

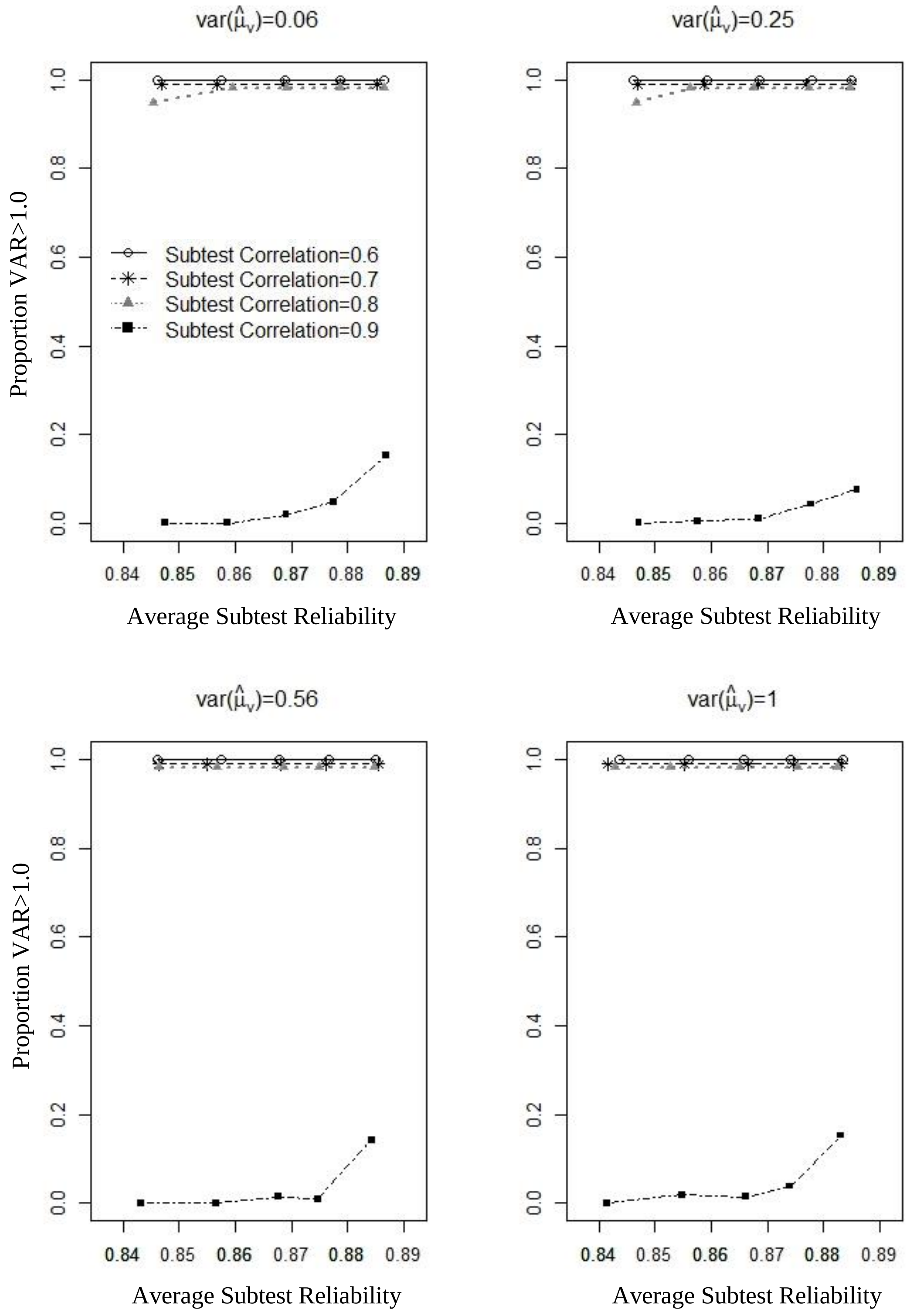


Figure 3. Scatterplot of $\mathcal{G}$ and VAR for Study A. Each marker corresponds to one of the 20 experimental conditions (five levels of reliability x four levels of correlation) within each level of $var(\bar{\mu}_v)$.

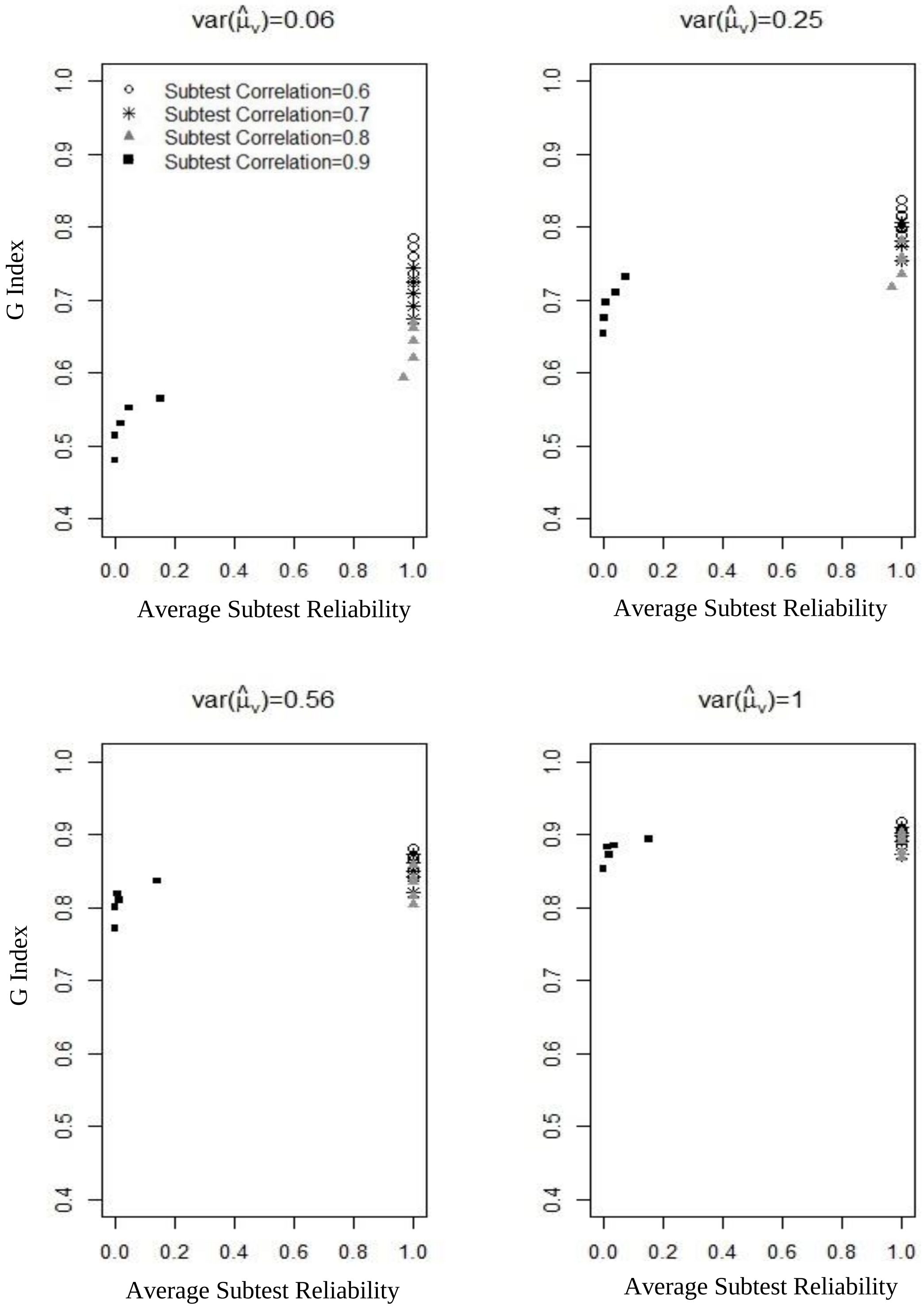
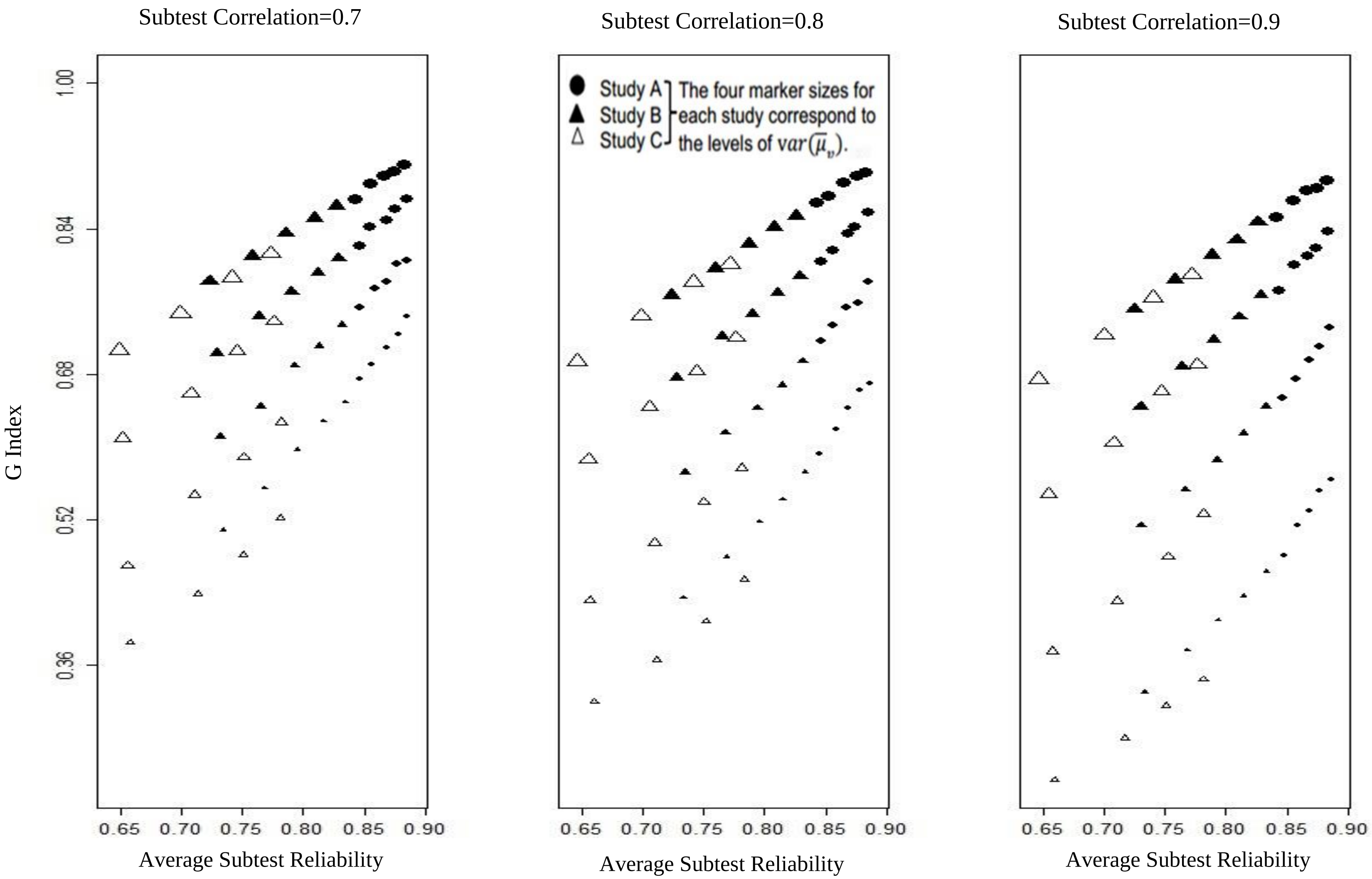


Figure 4. $\mathcal{G}$ as a function of subtest reliability for three levels of $\overline{\rho_{vv'}}$ across studies A, B, and C


Subtest Correlation=0.7
Subtest Correlation=0.8
Subtest Correlation=0.9
G Index
Average Subtest Reliability
● Study A
▲ Study B
△ Study C
The four marker sizes for each study correspond to the levels of $var(\overline{\mu}_v)$.