Using Multivariate Generalizability Theory to Evaluate Subscore Utility for Different Subgroups of Examinees

Zhehan Jiang – University of Kansas
Mark Raymond – National Board of Medical Examiners

Examinees and other test users often expect to receive subscores in addition to total test scores (Huff & Goodman, 2007). However, subscores have their limitations: often they are not empirically distinct from one another, and they tend to lack sufficient reliability to be of practical use. Indeed, the *Standards for Educational and Psychological Testing* (AERA et al., 2014) indicate that if interpretation of subscores, score differences, or profiles is suggested, then a rationale and relevant evidence in support of such interpretations should be provided (p. 27).

Various methods have been proposed for evaluating the usefulness of subscores, including visual inspection of correlations among subscores, factor analysis, and structural equation modeling (e.g., D'Agostino, Karpinski, & Welsh, 2011; Haladyna & Kramer, 2004; Stone, Ye, Zhu & Lane, 2010; Thissen, Wainer, & Wang, 1994). While these methods provide useful information, using the results to make decisions about subscore reporting involves an element of subjectivity. A method developed by Haberman (2008) removes the subjectivity. His approach incorporates both subscore distinctiveness and subscore reliability into a single objective decision rule. Brennan (2011), as well as Fienberg and Wainer (2014), have derived variations on Haberman's (2008) approach. Each of these methods is based on the principle that an observed subscore, V, is meaningful only if it can predict the true subscore, V_T, more accurately than the true subscore can be predicted from the total score Z. With Haberman's method, V_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 a V_T from V exceeds the PRMSE based on the total score Z, then the subscore adds value. In other words, subscores are useful only if observed subscores predict true subscores more accurately than total scores predict true subscores.

A consistent finding from numerous studies using PRMSE and other correlation-based methods is that subscores are seldom worth reporting (Puhan, Sinharay, Haberman, & Larkin, 2008; Sinharay, 2010; 2013; Stone et al., 2012). 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 of most tests often lack empirical support (Haberman et al., 2008; Harris & Hanson, 1991). Studies also have shown that, with a few exceptions, these conclusions are usually invariant with respect to subgroups based on gender, ethnicity, or other factors (Sinharay & Haberman, 2014). That is, empirical

findings suggest that if subscores are not useful for the total group, then they probably will not be useful for subgroups.

Although correlational methods are useful for summarizing relationships among variables, they have certain limitations. First, correlations overlook differences in means and variances across variables, and these sources of variation can be important when interpreting score profiles (Cronbach & Gleser, 1953). A second limitation is that correlations are relatively insensitive to substantial changes in score profiles subgroups of examinees. 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 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 $r_{xy} = .873$. That is, the change in correlation by adding one-half SD to half of the scores is only .027. Of course, the subgroup correlations remain at $r_{xy} = .90$, a consequence of adding a constant to all scores in the subgroup getting the intervention.[1] Although this example is contrived (an intervention never results in a constant effect), the empirical outcome is not: correlations are not sensitive to large systematic differences in subtest score distributions. There is at least one real-world counterpart to this example involving gender differences on essay tests. It is a general finding that females score higher than males on essay questions and lower on multiple-choice questions. However, correlations suggest that essay scores should not be separately reported (e.g., Thissen et al, 1994; Bridgeman and Lewis, 1994). Bridgeman and Lewis (1994) noted that exclusive reliance on correlations in this instance results in overlooking potentially important performance differences between men and women. It would seem that an evaluation of the utility of subscores would also consider the variability of subgroup score profiles.

An Index for the Reliability of Score Profiles

Cronbach and colleagues (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) laid a foundation for quantifying the properties of score profiles toward the end of the classic book, *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* Building on that foundation, Brennan (2001) introduced a reliability-like index for score profiles as part of his treatment of multivariate generalizability theory (G-theory). Brennan's index for score profile reliability, $\mathcal{G}$, indicates the proportion of variance in observed score profile variance attributable to universe (or true) score profile variance (Brennan, 2001, p. 323). Notably, $\mathcal{G}$ accounts for mean differences in score profiles.

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 $p \times (i{:}v)$. The multivariate designation of this design is $p\bullet \times i^{\circ}$, where the circles describe the multivariate

[1] This example is based on a sample of 100,000 simulated examinees. A similar effect is observed for correlations in the .70 to .90 range and with smaller samples of examinees.

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. For purpose of comparison, the *p• x i•* design, where both circles are solid, indicates that items are also linked. One example would be when examinees to each of several essays are rated on different attributes (e.g., grammar, organization). The *p• x i•* design can account for correlated error, while the *p• x i°* does not.

A multivariate G study based on the *p• 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) defines the generalizability index for score profiles as:

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

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 terms in numerator are:

$\overline{\sigma_v^2}(p)$ = 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 the 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• x i°* design, the covariance components for observed scores provides an unbiased estimate of covariance components for universe scores. That is, $\sigma_{vv'} = S_{vv'}$. The first term in both the numerator and the denominator, if considered alone, represents the ratio of true score variance to observed variance. Thus, it is apparent that $\mathcal{G}$ is essentially a reliability coefficient adjusted for covariances. 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 $\mathcal{G}$. It also is evident that

differences in subtest difficulty contribute to $\mathcal{G}$. If subscores all have the same mean for one group but different means for a second group, then $\mathcal{G}$ will be higher for the latter group all other things being equal.

To date, little research has been done on the potential use of G-Theory to evaluate the quality of subscores. One exploratory study reported that while $\mathcal{G}$ has moderately related to PRMSE, there were instances where PRMSE indicated that subscores are worth reporting, but $\mathcal{G}$ indices were in the 60s and 70s (Jiang & Raymond, 2017). In other instances, $\mathcal{G}$ indices were in the .80s, but PRMSE indicated that subscores did not add value. The most notable outcome was that differences in subtest means contributed substantially to $\mathcal{G}$, suggesting that it might be detect subscore differences in the utility of score profiles for different subgroups even in instances where subgroup correlations are similar and high.

The purpose of the present study is to evaluate sensitivity of $\mathcal{G}$ to differences in score profiles for subgroups of examinees. To provide a context for interpreting $\mathcal{G}$, we compare it to PRMSE. We simulate data for the two subtest case to facilitate that comparison. Since PRMSE evaluates each variable and $\mathcal{G}$ evaluates the entire score profile, the two become more difficult to compare for more than two subtests. Also, the two subtest case corresponds to a common data interpretation challenge in testing: whether to report subscores when a test consists of both standard multiple-choice questions (MCQs) and some other format such as constructed response science items or essays (Bridgeman & Lewis, 1994; Bridgeman, 2016).

We conduct a simulation study to evaluate the properties of $\mathcal{G}$ over various conditions. For each condition, there was a *reference group* whose score profile was flat, and a lower performing *focal group* whose score profile varied. Because PRMSE is generally known, we use it as a basis for comparison – not so much for determining superiority but as a means for understanding the properties of $\mathcal{G}$ and when it might be useful. The study addresses the following questions: (1) To what extent are $\mathcal{G}$ and VAR differentially sensitive to differences in subtest means for subgroups of examinees? (2) To what extent do those differences vary by subtest reliabilities and correlations? (3) What conditions lead to different decisions for $\mathcal{G}$ and PRMSE?

Method

Overview

We conduct a series of simulations to evaluate the properties of $\mathcal{G}$ and PRMSE over various conditions of total test and subtest reliability, subtest correlations, and variation in subtest means. For each condition, there was a *reference group* whose score profiles were flat, and a lower performing *focal group* whose score profiles varied. The first two factors (reliability, correlations) have been studied in prior simulations of subscore utility, while the last (variation in means) has not. Total test and subtest reliability were controlled as a single unit because of their obvious dependence.

Independent Variables

We simulate item responses for two groups of examinees: a reference group, and a lower performing focal group. Within each group the three factors are investigated:

- Population correlation, $\rho_{vv\prime}$, between subtests. Three levels were studied, with values of $\rho_{vv\prime}$ = .73, .81, and .90. These values are comparable to the subtest correlations often seen in the literature (e.g., Sinharay, 2010; Sinharay & Haberman, 2014).
- Subtest reliability, ρ_v^2. Three conditions 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 levels of reliability. Note that total test reliability naturally covaried with subtest reliability (.94, .90, and .82). However total test reliability has no direct effect on $\mathcal{G}$ and, while useful to keep in mind, is of little consequence for this study.
- Difference in population subtest means, $\Delta\underline{\mu}$, for the focal group: Four levels of $\Delta\mu$ were controlled to be 0.00, 0.25, 0.50, and 0.75. Figure 1 depicts the simulated means for the two subtests at the four levels of $\Delta\underline{\mu}$. These differences are comparable to some of the values reported in practice (e.g., Bridgeman & Lewis, 1994; Sinharay & Haberman, 2014), although ΔM = 0.75 is on the high side. The reference groups always had no variation in means ($\underline{\mu}$ = 1.0 for both subtests in all conditions). The overall difference in means between the reference and focal groups is of no consequence as both $\mathcal{G}$ and PRMSE were computed within groups.

Experimental conditions were created by crossing these three factors within the reference group and focal group. Thus, there were 72 total conditions, 36 within each group. The number of subtests was fixed at two for the previously noted reasons and because other studies of subscore utility suggest that results generalize from the two subtest case to multiple subtests (Feinberg & Wainer, 2014; Jiang & Raymond, 2017). We set sample size at n = 1000 per group. Sample size was not manipulated because previous simulations on the utility of subscores indicate that when samples sizes are reasonably large it has little impact on results.

Item Response Simulation

Subscores for simulated examinees 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 parameters were generated from a log-normal distribution (*M* = 0.0, *SD* = 0.5), while difficulty parameters were normally distributed (*M* = 0, *SD* = 1). 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. Four mean vectors of ability parameters $\boldsymbol{\mu}$ were specified for the focal group to produce to the values of $\Delta\underline{\mu}$ 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 $\rho_{vv\prime}$ and was assigned values of .73, .81, and .90.

Outcome Variables

The two outcomes of interest are $\mathcal{G}$ and proportion reduction in mean-squared error (PRMSE) (Haberman, 2008). The goal is not so much to compare the merits of $\mathcal{G}$ to those of PRMSE, but rather to use the latter as a baseline for comparison and interpretation of $\mathcal{G}$, since the measurement community has experience with PRMSE. Brennan's (2001) $\mathcal{G}$ 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). We also 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 particular replication. Both $\mathcal{G}$ and VAR were computed for each replication. For each of the 36 conditions within each group (focal, reference), we report the mean $\mathcal{G}$ across the 200 replications. As VAR is a dichotomous decision (0, 1), we report the proportion of the 200 replications for which VAR ≥1.

Results

Figure 2 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 indicate the three levels of subtest correlation ($\rho_{vv\prime}$= .73, .81, .90). Keep in mind that the reference group always had $\Delta\mu = 0$, thus the x-axis displays only one level for the focus group. We first consider results for $\mathcal{G}$ and VAR separately, and then discuss the two indices together.

Across all conditions, $\mathcal{G}$ ranged from .27 to .82, with an overall mean of .53. However, values of $\mathcal{G}$ for the reference group were consistently lower than for the focal group (mean $\mathcal{G}$ = .48 and .58, respectively. Within any single panel in the left portion of Figure 2, $\mathcal{G}$ increased as the subtest correlations

dropped and as variation in subtest means increased. Looking down across the panels, it can be seen that $\mathcal{G}$ declines with lower levels of subtest reliability. Overall, the values of $\mathcal{G}$ generally are quite modest, even for conditions where one might expect it to be high. For example, under the conditions most favorable to subscores (panel A, top line), $\mathcal{G}$ = .70 for the reference group and ranged from .72 to .82 for the focal group. Under the least favorable conditions (panel C, bottom line), $\mathcal{G}$ = .28 for the reference group and ranged from only .29 to .52 for the focal group. In all conditions where focal group score profiles were not flat (i.e., $\Delta\mu > 0$), their $\mathcal{G}$ index was higher than for the reference group.

The panels on the right side of Figure 2 show the proportion of times VAR was greater than 1.0 for each condition. When interpreting VAR for a single replication, a value of 1.0 is required to conclude that subscores are worth reporting. However, when aggregating VAR indices across multiple replications for a particular condition, a overall threshold of .50 is reasonable, because VAR would exceed 1.0 more often than not.

The mean VAR across all conditions and all three panels was .48. As Figure 1 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 reliability (ρ_v^2 = .83, .89), and when subtest correlations are not excessively high ($\rho_{vv\prime}$ =.73, .81). These findings are not unlike those reported by Sinharay (2010). Notably, the decision of whether to report subscores based on VAR was usually the same for the reference group and focal group. The exception is for the conditions represented in the center panel where ρ_v^2= .83 and $\rho_{vv\prime}$=.81; in this condition VAR exhibited some sensitivity to the differences in the focal group.

Although $\mathcal{G}$ and VAR clearly differ, Figure 2 suggests that they covary. To clarify their relationship, Figure 3 presents a scatterplot between $\mathcal{G}$ and VAR. The open circles correspond to the reference group, while the triangles correspond to the focal group. One notable observation is that for all conditions where $\Delta\mu$ = 0, VAR indicates that subscores are worth reporting for all values of $\mathcal{G}$ that exceed .60. That is, some relatively unreliable score profiles were deemed worth reporting according to VAR. A second observation is that for all conditions where $\Delta\mu$ = 0, which includes all 36 reference group conditions (open circles) and 9 focal group conditions (inverted triangles), the relationship between the VAR and $\mathcal{G}$ 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. However, for focal group conditions where $\Delta\mu > 0$ (upright triangles), $\mathcal{G}$ increases relative to VAR and the two indices diverge from the logistic function. In these instances $\mathcal{G}$ looks better than what one might expect based on VAR. The largest outlier occurred for the condition where ρ_v^2= .89, $\rho_{vv\prime}$=.90, and $\Delta\mu$ = .75; at the condition VAR =.44 while $\mathcal{G}$ =.72.

Differences between VAR and $\mathcal{G}$ can be further illustrated by referring to Figure 2. Assume for the moment that a threshold of $\mathcal{G} > .70$ has been established to allow subscores to be reported. Now consider

the top-right and top-left panels where subtests are very reliable (ρ_v^2 = .89) and the two lines where subtest correlations are not excessively high ($\rho_{vv'}$ =.73, .81). VAR deems subscores to be worth reporting for *all* conditions for both the reference group and the focal group; in other words, VAR is invariant with respect to group membership for these conditions. However, $\mathcal{G}$ paints a different landscape. Under the decision rule adopted above ($\mathcal{G}$ = .70), one would conclude that subscores for the reference group are not worth reporting; however, subscores for the focal group would be reported for six of eight conditions. Lowering the threshold to $\mathcal{G}$ = .65 slightly improves agreement with VAR. It would allow subscores for the reference group at $\rho_{vv'}$ =.73, but not at $\rho_{vv'}$= .81. Meanwhile, subscores for the focal group would be judged as reportable for all of the conditions being discussed if the threshold were set at .65. In short, VAR appears to be tolerant of score profiles for which the proportion of true score variance as captured in $\mathcal{G}$ is less than optimal.

Discussion

Previous work evaluating the utility of score profiles for subgroups of examinees has relied almost exclusively on subgroup correlations and reliability coefficients as the basis for evaluation, even though subgroups may have score profiles that exhibit considerable variability in subtest means (see Sinharay & Haberman, 2014 for examples). The present study examined the sensitivity to subgroup differences of Brennan's (2001) $\mathcal{G}$, a reliability-like index suitable for score profiles. Given that the equation for $\mathcal{G}$ includes terms not only for reliability and correlations, but also for subscore means, it seems particularly appropriate for circumstances where score profiles for one or more subgroups of examinees are not flat. The simulations reported here produced several findings that shed light on the potential utility of $\mathcal{G}$. As expected, $\mathcal{G}$ increased with higher levels of subtest reliability, greater differences in subtest means, and lower levels of subtest correlation. However, values of $\mathcal{G}$ seemed almost surprisingly low, seldom reaching what one might regard as 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. These particular findings are consistent with those reported in an earlier study (Jiang and Raymond, 2017). The result of primary interest for this study pertains to subgroup differences, where $\mathcal{G}$ was found to be sensitive to group differences in subscore means. More specifically, $\mathcal{G}$ was consistently higher for the focal group where subtest means varied, than for the reference group where subtest means were equal.

The data also indicated that VAR and $\mathcal{G}$ are closely related under certain conditions. For those instances where 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, $\mathcal{G}$ increased and the logistic function no longer described its relationship with VAR. As indicated in Figure 3, subscores with $\mathcal{G}$ coefficients as low as the .60s were deemed by VAR as

reportable. This result is a little concerning because it indicates that VAR can be quite tolerant of score profiles with a fair amount of measurement error. In other words, subscores deemed worthy of reporting according to VAR may not meet generally acceptable levels of profile reliability. Under conditions for which subtest means are equal, $\mathcal{G}$ may be seen as more conservative than VAR.

In conclusion, the findings indicate that $\mathcal{G}$ is a useful adjunct to PRMSE or VAR for evaluating subscore utility. There were instances for which $\mathcal{G}$ was more sensitive to group differences on subtests, which can be important to consider when deciding to report subscores where different subgroups may have different mean score profiles (e.g., women scoring better on essay-based subtests; students receiving different interventions). Further study of $\mathcal{G}$ seems warranted for a few reasons. First, this study was modest in size and scope; its limitations should be addressed in future investigations. For example, it would be important to extend the conditions studied here (different levels of ρ_v^2 , $\rho_{vv\prime}$, and $\Delta\mu$), and to manipulate other conditions (e.g., sample size, percent of sample in focal group). Second, the present study used VAR > 1.0 as a dichotomous outcome (report subscores or not), which is consistent with best practices. However, future efforts should just look at the mean levels of VAR to better understand its relationship with $\mathcal{G}$. Finally, the measurement community would benefit from guidelines on how to interpret $\mathcal{G}$. While $\mathcal{G}$ indices in the .70s certainly seem low, further study could suggest otherwise… or not. Guidelines for interpretation typically evolve over time as researchers gain experience with a procedure. It is hoped that this initial effort encourages further research into the properties of $\mathcal{G}$.

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.

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.

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.

Huff, K., & Goodman, D.P. (2007). The demand for cognitive diagnostic 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.

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.

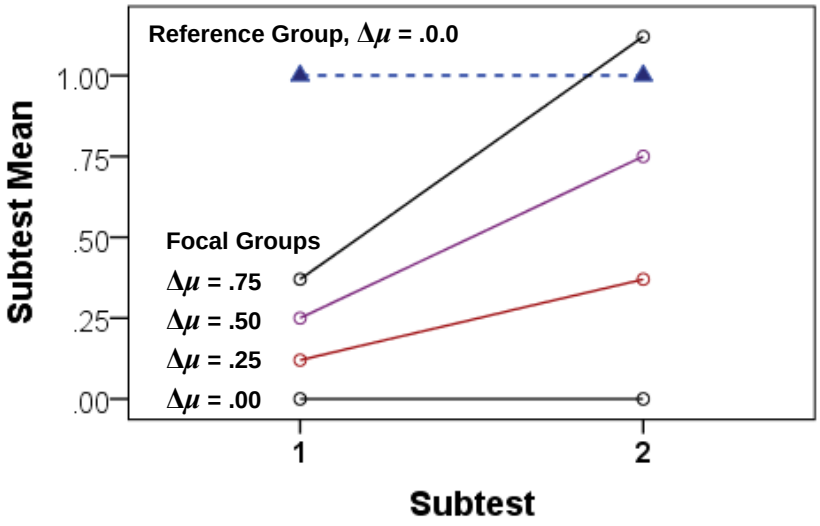

Figure 1. Subtest population means, and differences in subtest means ($\Delta\mu$), for reference and focal groups for simulated data.

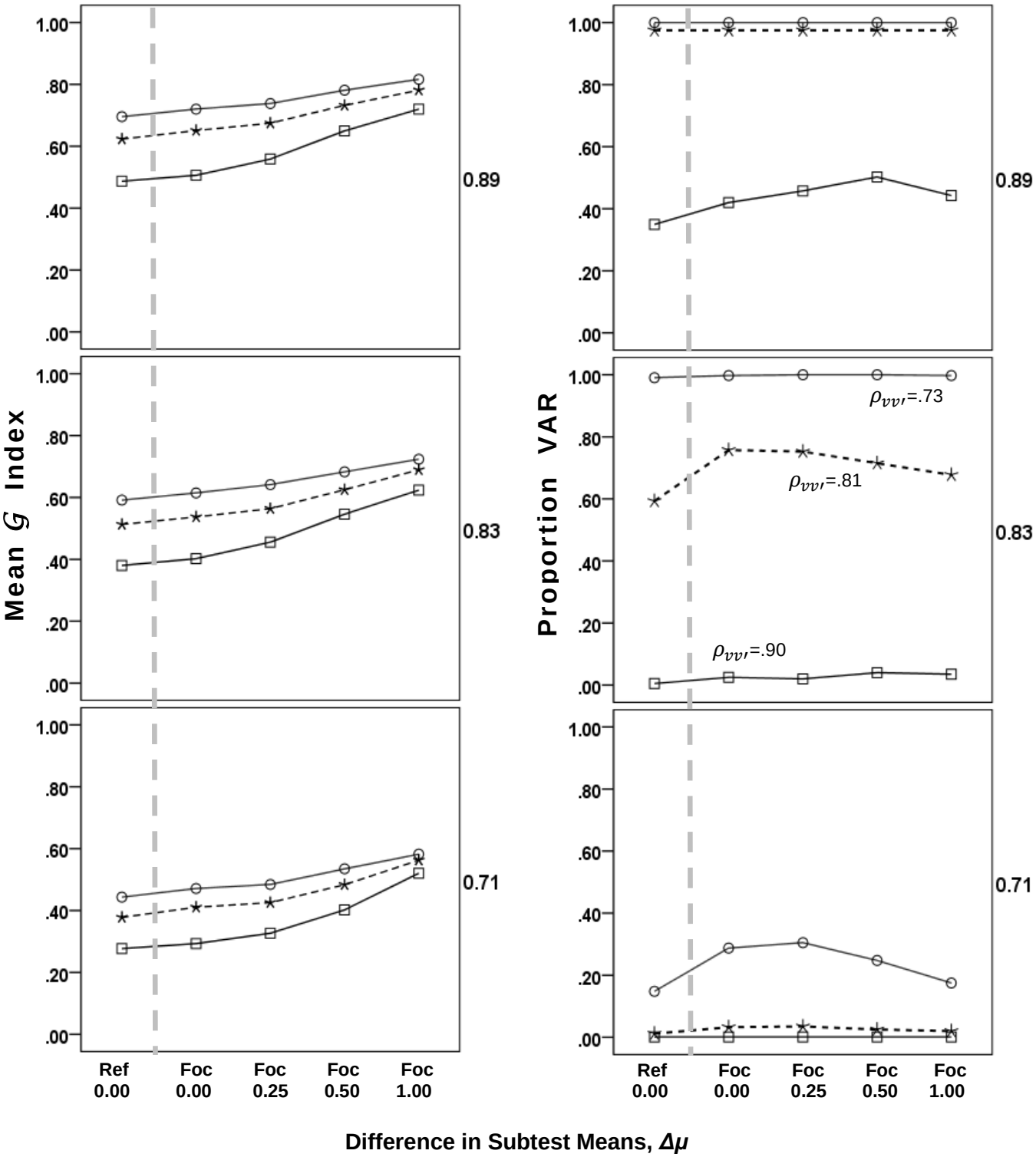

Figure 2: Mean $\mathcal{G}$ and proportion VAR as a function of differences in subtest means for the focal group. There are three levels of subtest reliability (ρ_v^2= .71, .83, .89) presented in the three vertical panels, and three levels of true score correlation ($\rho_{vv\prime}$= .73, .81, .90) represented by the three lines. Note that the reference subtest means were always equal ($\Delta\mu$ = 0.00).

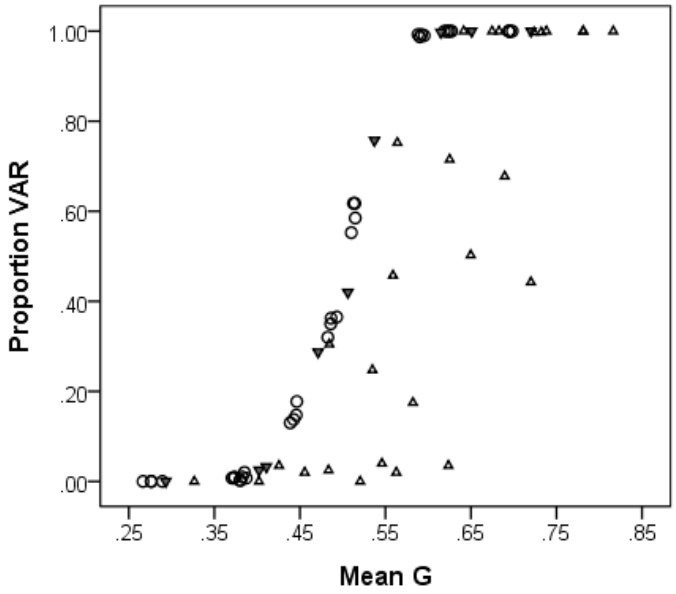

Figure 3. Relationship between proportion VAR and mean $\mathcal{G}$ for all experimental conditions. Open circles indicate the reference group ($\Delta\mu = 0$); inverted triangles indicate the focal group where $\Delta\mu = 0$; while the upright triangles indicate the focal group where $\Delta\mu > 0$. 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 ($R^2 = .95$).