Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 59

Reliability of Difference Scores Obtained from Nested Data within a Multivariate Generalizability Theory Framework

Rabia Karatoprak Ersen[†]
Won-Chan Lee
Donald B. Yarbrough

August 2025

[†]Rabia Karatoprak Ersen is Postdoctoral Researcher, Department of Survey Data Curation, GESIS – Leibniz Institute for the Social Sciences (email: Rabia.KaratoprakErsen@gesis.org). Won-Chan Lee is Director, Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa (email: won-chan-lee@uiowa.edu). Donald B. Yarbrough is Professor Emeritus of Educational Measurement and Statistics, Department of Psychological and Quantitative Foundations, University of Iowa (email: d-yarbrough@uiowa.edu)

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Contents

List of Tables

List of Figures

Abstract

The purpose of this study is to examine the reliability and dependability of difference scores computed as the change between a pretest and a posttest administered to assess the effectiveness of an intervention. The data collection design involved a nested structure, with persons (p) nested within groups (g), and groups nested within sites (s). Multivariate generalizability theory was used to analyze the data to evaluate the reliability and dependability of difference scores at the levels of persons, groups, and sites. The G study designs included: $(p^\bullet : g^\bullet : s^\bullet) \times i^\bullet$, $(p^\bullet : s^\bullet) \times i^\bullet$, $(p^\bullet : g^\bullet) \times i^\bullet$, $g^\bullet \times i^\bullet$, $p^\bullet \times i^\bullet$, and $s^\bullet \times i^\bullet$ with pretest and posttest serving as the two levels of the multivariate facet. In the designs with sites as the object of measurement: $(P^\bullet : G^\bullet : s^\bullet) \times I^\bullet$, $(P^\bullet : s^\bullet) \times I^\bullet$, and $s^\bullet \times I^\bullet$, omitting groups within sites or persons within groups in the design led to underestimation of error variances and inflated generalizability and dependability coefficients. The relative error correlations increased, and the absolute error correlations decreased across $s^\bullet \times I^\bullet$, $(P^\bullet : s^\bullet) \times I^\bullet$, and $(P^\bullet : G^\bullet : s^\bullet) \times I^\bullet$. Across all designs, generalizability and dependability coefficients were similar in magnitude, primarily due to the relatively small item variance. Compared across different object of measurements, both the generalizability and dependability coefficients were highest when the object of measurement was persons, and lowest when it was sites.

1 Introduction

In practice, difference scores are utilized for various purposes under longitudinal or repeated measurement designs. For instance, experimental research designs are conducted to assess the effectiveness of new instruction methods, educational programs, or treatments. This involves administering pretests and posttests in multiple schools. Another example of interest concerns assessing changes in academic success. For instance, the Every Student Succeeds Act requires states to measure academic progress in elementary and middle schools. Thus, schools are responsible for demonstrating that they provide high-quality education. One of the indicators is academic progress. Then administrators want to know how much academic progress students, classes, and thus schools achieved between different grades. In these cases, inferences can be made using difference scores computed by subtracting pretest scores from posttest scores. And the data from which the difference scores are obtained, has nested structure such that students are nested within classrooms and classrooms are nested within schools.

The Standards for Educational and Psychological Testing (2014) state that whenever the difference between observed scores is interpreted and used for actions, reliability of difference scores needs to be reported. Moreover, if there is an intent or need to make interpretations at the group level, then reliability of group means also needs to be reported. When difference scores are computed at the group level, reliability of group mean difference scores are required to be reported. Generalizability theory (G theory) provides extensive tools for estimating reliability including reliability of difference scores and group scores while considering various sources of measurement error (Brennan, 2001a). The reporting of group mean difference scores has been examined in multiple studies using classical test theory (CTT) or G theory (e.g., Brennan, 1995; Cronbach et al, 1997, Brennan et al., 2003). However, the data in these studies did not encapsulate the nested structure of schools, which changes the G theory study design. Moreover, they did not study the effect of ignoring lower-level variances on reliability of school mean difference scores.

The previous studies focusing on the reliability of group mean difference scores (Yin & Brennan, 2002; Brennan, Yin, & Kane, 2003) did not use data that represent nested structure of the schools. Consequently, the variation in the object of measurement within a single dataset was not studied and the importance of incorporating nesting when calculating the reliability of scores derived from a specific object of measurement was overlooked. The purpose of this study is to fill this gap by demonstrating what the sources of variance are, how they contribute to the reliability, and how they change when the object of measurement changes in a nested design.

Wei and Haertel (2011) conducted a study on the reliability of school means using univariate G theory and concluded that classroom-level variance needs to be included in estimating reliability. They asserted that excluding classroom-level variance can introduce bias into the reliability estimates. Reliability of difference scores requires consideration of correlated error between the pretest and posttest from which difference scores are computed. Multivariate G theory

is pertinent in such cases such that variance covariance components can be explicitly modeled in computation of correlated errors. The second purpose of this study is to examine the effect of ignoring levels of nested data in the computation of reliability for difference scores.

For these purposes, this study used an empirical data set consisting of pretest and posttest data administered to assess the effectiveness of a statewide intervention (Wade & Yarbrough, 2007). Teams of teachers from various school districts implemented the intervention for students. Within each school district, multiple teachers implemented the intervention, and each teacher worked with multiple students. Thus, the data had a nested structure such that students were nested within teachers and teachers were nested within school districts. In this study, the terms persons (p), groups (g), and sites (s) are used to refer to students, teachers, and school districts, respectively. Prior research on difference scores often examines students nested within classes and classes nested within schools. However, in this study, groups are equivalent to classes (one level of nesting), and sites are equivalent to schools (two levels of nesting). With this framework, this study aims to answer the following research questions:

1. How does the reliability of difference scores compare when the objects of measurement are (a) sites, (b) groups, or (c) persons?
2. What are the impacts of neglecting nested facets, such as persons or groups, on the reliability of difference scores?

The remainder of this paper begins with a brief overview of CTT and G theory, followed by a review of previous research on difference scores. Then, the study design and analysis procedures are described. It concludes with a discussion of results in relation to the research questions.

1.1 CTT, Univariate G theory, Multivariate G theory

Suppose individuals respond to a set of essay prompts, and raters evaluate their responses. In this assessment scenario, essay prompts and raters are key sources of measurement error that should be incorporated into reliability estimation. CTT-based reliability estimation methods treat error as a single undifferentiated entity and cannot separate multiple sources of error in the estimation process. However, G theory offers methodologies capable of accounting for different sources of error, allowing for the examination of their individual contributions to reliability.

Furthermore, G theory explicitly distinguishes between two types of error: relative error (δ) and absolute error (Δ). Relative error variance ($\sigma^2(\delta)$) is used to estimate the generalizability coefficient ($\mathbf{E}\hat{\rho}^2$), which serves as an index for norm-referenced interpretations. Absolute error variance ($\sigma^2(\Delta)$) is used to estimate the dependability coefficient ($\hat{\Phi}$), which is an index for domain-referenced inferences. These error types also serve as the basis for other indices such as the signal-noise ratio (S/N; Brennan & Kane, 1977) and the error-tolerance ratio (E/T; Kane, 1996). In contrast, reliability indices in CTT involve relative error

variance, making them more suitable for norm-referenced interpretations. CTT typically does not distinguish between absolute and relative error.

G theory constitutes both univariate and multivariate methodologies. One of the distinct advantages of multivariate G theory over univariate G theory is its ability to account for correlated error variances and to model composite scores. For instance, difference scores, such as the difference between posttest and pretest scores, can be specified as composite scores. In this framework, both variance and covariance components for pretest and posttest scores can be estimated and used in the calculation of reliability coefficients for difference scores. However, univariate G theory approach does not have the capacity to model difference scores as composites or to estimate covariance between pretest and posttest scores. Typically, CTT assumes zero covariance between errors of pretest and posttest scores. Multivariate G theory relaxes these assumptions, allowing for a more accurate and flexible estimation of the reliability of difference scores.

1.2 Research on Difference Scores

Researchers often avoid using difference scores mainly for several reasons, as outlined by Gu et al. (2018). First, difference scores tend to have low reliability, implying a weak correlation between observed and true difference scores. Second, the reliability of difference scores is generally lower than that of the test scores from which the difference scores are obtained. Third, when reliability is low, the measurement precision of difference scores at the individual level becomes limited. Fourth, both a high correlation between pretest and posttest scores and a high reliability of difference scores cannot be observed at the same time. Finally, difference scores can be potentially defective, given that the correlation between difference scores and pretest scores can be negative, raising concerns about their interpretability.

However, studies have provided evidence contradicting the notion that psychometric properties of difference scores are inadequate for making valid inferences. Gu et al. (2018) provided analytical explanations based on CTT by differentiating individual and group level change scores. Moreover, difference scores can be useful if indices for reliability are selected according to the intended interpretations which may be norm-referenced or domain-referenced (Yin & Brennan, 2002). G theory can provide $\mathbf{E}\hat{\rho}^2$, $\hat{\Phi}$, S/N, or E/T for both norm- and domain-referenced interpretations (Brennan & Kane, 1977; Brennan et al., 2003; Kane, 1996; Miller & Kane, 2001).

A key determinant for which reliability index should be used is the intended interpretations of difference scores. If the focus is not on ranking individuals based on their difference scores, the appropriate index to use is a coefficient representing the reliability of absolute change, for which G theory is particularly well suited (Yin & Brennan, 2002). Miller and Kane (2001) proposed using E/T, which can be computed using $\sigma^2(\delta)$ or $\sigma^2(\Delta)$ for norm-referenced or domain-referenced interpretations. Brennan et al. (2003) examined the reliability of group mean difference scores using both $\mathbf{E}\hat{\rho}^2$ and $\hat{\Phi}$ as well as the

E/T ratio. Their analysis, conducted using multivariate G theory, revealed that both relative-type and absolute-type errors were correlated in their data.

Existing studies using the G theory framework differ in terms of their G study designs. Yin and Brennan (2002) did not strictly adhere to G theory conventions, but their design can be reasonably interpreted as a $p \times i$ design. The study design of Miller and Kane (2001) was $p \times (i : c)$, where persons were crossed with items and items were nested within categories. Brennan et al. (2003) used $p : (s \times c)$ where persons were nested within school districts and districts were crossed with cohorts. In their multivariate G theory analysis, this design was extended to $p^\bullet : (s^\bullet \times c^\bullet)$, where the object of measurement was groups rather than individuals. The inclusion of cohorts served to replicate the measurement.

When estimating reliability of difference scores for groups, the nested structure of the data is a source of variance. In Yin and Brennan (2002), difference scores for sites (i.e., school districts) were examined using CTT. In Brennan et al. (2003), difference scores were expressed in grade equivalents for a site, using a $p : (s \times c)$ design. However, despite the nested nature of the data, where persons are within groups and groups within sites, this nesting was not incorporated into the reliability analysis. As Wei and Haertel (2011) noted, ignoring group level variance can introduce bias in reliability estimates for site scores. The purpose of this study is to address this gap by examining reliability of difference scores when data involve a two-level nested structure, using multivariate G theory.

2 Methods

2.1 Data

This study used an empirical data set obtained from a pretest and a posttest administered to persons nested within groups, which were in turn nested within sites. The following notational conventions are used: p for persons (i.e., students), g for groups, s for sites, and i for items. The same instrument was used for both administrations and consisted of 25 Likert-type items (n_i=25), each with four response categories. The total sample included 1,517 persons (n_p=1,517). The number of persons within groups ranged from 2 to 62. There were 89 groups (n_g=89) across 32 sites (n_s=32). The number of groups per site ranged from 1 to 8.

2.2 Designs

In this study, multiple analyses were conducted using multivariate G theory. The reporting starts with the $(p^\bullet : g^\bullet : s^\bullet) \times i^\bullet$ design, which represents the full multilevel structure of the data. Additional G study designs were $(p^\bullet : s^\bullet) \times i^\bullet$, $(p^\bullet : g^\bullet) \times i^\bullet$, $p^\bullet \times i^\bullet$, $g^\bullet \times i^\bullet$, and $s^\bullet \times i^\bullet$. Site was an aggregate unit of analysis with two levels of nesting, whereas group represented an aggregate

unit of analysis with one level of nesting. Persons, groups, sites, and items were treated as random facets in both the universe of admissible observations and the universe of generalization.

All these designs were multivariate G study designs, with difference scores treated as composite scores–i.e., $x_{diff} = x_{post} - x_{pre}$. The fixed multivariate variable (ν) had two levels: pretest and posttest. Thus, the weights (w) for pretest and posttest were -1 and $+1$, respectively. In the notation used, a circle as a superscript on a facet designates its relationship to the fixed multivariate variable. A filled circle denotes that the facet is crossed with the multivariate levels, while an empty circle indicates that the facet is nested within the multivariate levels. In this study, the multivariate designs involved filled circles for all facets, as pretest and posttest scores were available for all persons within all groups and sites. This resulted in full and symmetric variance-covariance matrices.

Analysis of all G studies except the $(p^\bullet : g^\bullet : s^\bullet) \times i^\bullet$ were conducted using mGENOVA (Brennan, 2001b). Since mGENOVA does not support the estimation of variance and covariance components for $(p^\bullet : g^\bullet : s^\bullet) \times i^\bullet$, urGENOVA (Brennan, 2001) and supplemental hand-computations were used for that analysis. Further details are provided in the corresponding section.

For all G studies, a variance-covariance matrix was generated for each score effect α, universe score, relative error, and absolute error. Table 1 illustrates the components used in the estimation of $\sigma^2(\tau)$, $\sigma^2(\delta)$, and $\sigma^2(\Delta)$ for each design.

Composite universe score variance, $\sigma^2_{diff}(\tau)$, composite error variances, $\sigma^2_{diff}(\delta)$ and $\sigma^2_{diff}(\Delta)$, were computed using the following equation:

$$\sigma^2_{diff}(\alpha) = \sigma^2_{post}(\alpha) + \sigma^2_{pre}(\alpha) - 2\sigma_{pre,post}(\alpha). \tag{1}$$

Generalizability coefficient is:

$$\mathbf{E}\rho^2 = \frac{\sigma^2_{diff}(\tau)}{\sigma^2_{diff}(\tau) + \sigma^2_{diff}(\delta)}. \tag{2}$$

Dependability coefficient is:

$$\Phi = \frac{\hat{\sigma}^2_{diff}(\tau)}{\hat{\sigma}^2_{diff}(\tau) + \hat{\sigma}^2_{diff}(\Delta)}. \tag{3}$$

2.2.1 $(p^\bullet : g^\bullet : s^\bullet) \times i^\bullet$ and $(P^\bullet : G^\bullet : s^\bullet) \times I^\bullet$

Difference scores at the site level, where persons are nested within groups and groups are nested within sites, are modeled using $(p^\bullet : g^\bullet : s^\bullet) \times i^\bullet$ design, as illustrated in Figure 1. urGENOVA (Brennan, 2001) can compute variance components for the univariate $(p : g : s) \times i$ design, but it does not provide covariance estimates between pretest and posttest scores for the multivariate $(p^\bullet : g^\bullet : s^\bullet) \times i^\bullet$ design. To address this limitation, covariance estimates for $(p^\bullet : g^\bullet : s^\bullet) \times i^\bullet$ were obtained by combining urGENOVA outputs with the variance-of-a-sum procedure (Brennan, 2001a, p. 361). Specifically, variance

components $\sigma^2(\alpha)$ were first estimated using urGENOVA for the posttest scores (x_{post}), pretest scores (x_{pre}) and sum scores ($x_{sum} = x_{post} + x_{pre}$). Then, the covariance component $\sigma_{pre,post}(\alpha)$ for each effect was derived using the following equation:

$$\sigma_{pre,post}(\alpha) = \frac{\sigma^2_{sum}(\alpha) - \sigma^2_{pre}(\alpha) + \sigma^2_{post}(\alpha)}{2}. \quad (4)$$

Similarly, both relative and absolute error variances of x_{pre} and of x_{post} were obtained from urGENOVA. Error covariances between x_{pre} and x_{post}, for both relative and absolute error, were derived from Equation 4. Table 2 summarizes the estimated G study variance and covariance components for both pretest and posttest scores. The correlations reported in the table are disattenuated correlations, which were computed according to

$$\rho_{pre,post}(\alpha) = \frac{\sigma_{pre,post}(\alpha)}{\sqrt{\sigma^2_{pre}(\alpha)\sigma^2_{post}(\alpha)}}. \quad (5)$$

As shown in Table 2, the estimated variance component of posttest scores for sites ($\hat{\sigma}^2_{post}(s)$) was higher than that of pretest scores ($\hat{\sigma}^2_{pre}(s)$). Similarly, the posttest variance components for $g:s$, $p:g:s$, and si were larger than their corresponding pretest variance components. In contrast, the posttest variance components for i, $gi:s$, and $pi:g:s$ were smaller than those for pretest scores. Despite these differences in magnitude, the pattern of variance components was generally consistent across pretest and posttest scores. For instance, the variance component for $pi:g:s$, which represents the variance in observed scores for persons over items within a randomly selected group and site, was the largest estimated variance component. This was expected, as it captures the residual variance, encompassing all sources of variation that is not represented in the universe of admissible observations (Brennan, 2001). The second largest variance component was for $p:g:s$ which indicates substantial variability among persons nested within groups and sites. This result was expected, given the presence of two nesting facets, which would otherwise have contributed to crossed sources of variation. The estimated variance components for s ($\hat{\sigma}^2_{pre}(s)$) and si ($\hat{\sigma}^2_{pre}(si)$) in the pretest data were similar in magnitude and represented the smallest among all components.

As shown in Table 2, the estimated disattenuated correlation between pretest and posttest scores for items (i.e., $\hat{\rho}_{pre,post}(i)$) was very high, suggesting a strong linear relationship between item scores across the two occasions. The correlation for the site-by-item interaction ($\hat{\rho}_{pre,post}(si)$) was relatively high, suggesting that the rank ordering of sites by item remained largely consistent from pretest to posttest. By contrast, $\hat{\rho}_{pre,post}(s)$ was the lowest among all components except for the residual term $pi:g:s$. This suggests that the rank ordering of sites' pretest mean scores and posttest mean scores was not strongly preserved. Notably, $\hat{\sigma}^2_{post}(s)$ was approximately twice as large as $\hat{\sigma}^2_{pre}(s)$. These findings indicate that the intervention contributed to an increase in the variability of scores at the site level. Additionally, $\hat{\sigma}^2_{pre}(p:g:s)$ was almost ten times greater than $\hat{\sigma}^2_{pre}(g:s)$, and $\hat{\sigma}^2_{pre}(g:s)$ was almost twice as large as $\hat{\sigma}^2_{pre}(s)$.

Moreover, the difference between pretest and posttest variance components was larger for s than $g : s$. Similarly, the difference was larger for $g : s$ than $p : g : s$. This suggests that the intervention had a more substantial effect on aggregated levels of measurement.

Table 3 summarizes the variance and covariance components from the D study random effects design, denoted as $(P^\bullet : G^\bullet : s^\bullet) \times I^\bullet$, in which sites are the object of measurement. Table 3 includes G study sample sizes used as divisors in calculating the D study variance components. The number of persons within each group and site ($n_{p:g:s}$) was different across groups and sites, ranging from 2 to 62. The harmonic mean of $n_{p:g:s}$ was used as the divisor. The calculation was performed in two steps. First, for each site, the harmonic mean of the number of persons across groups was computed, yielding the average number of persons per group for that site. Then, the harmonic mean of these site-level harmonic means was computed. This process resulted in 8.277. There were a total of 89 groups and 32 sites in the sample. The number of groups within each site ($n_{g:s}$) ranged from 1 to 8. To determine the average number of groups per site, the harmonic mean of the number of groups across all sites was computed. This process resulted in 2.349. The number of items (n_i) was 25.

The universe of generalization in this study involved persons, groups, and items. This means that generalization was made from site scores obtained from specific items answered by specific persons within specific groups to the site scores which would be obtained from randomly sampled items, persons, and groups from infinitely large universes. In both the universe of generalization and the D study design, items were crossed with sites, while groups were nested within sites, and persons were nested within groups. This can be interpreted as sites' observed mean difference score was associated with approximately 2 groups per site, with approximately 8 persons within a group.

The universe score variance-covariance matrix was obtained as

$$\Sigma_\tau = \Sigma_s = \begin{bmatrix} \sigma^2_{pre}(s) & \sigma_{pre,post}(s) \\ \sigma_{pre,post}(s) & \sigma^2_{post}(s) \end{bmatrix}. \tag{6}$$

The estimated universe score variances for sites were $\hat{\sigma}^2_{pre}(\tau) = 0.013$ and $\hat{\sigma}^2_{post}(\tau) = 0.024$. The estimated correlation between site mean universe scores ($\hat{\rho}_{pre,post}(\tau) = 0.432$) was positive but relatively low. This suggests that the rank ordering of sites differed between pretest and posttest. This pattern is commonly observed with difference scores and is one of the key reasons they often exhibit low reliability.

The relative error variance-covariance matrix was calculated using the G study variance-covariance matrices as

$$\begin{aligned} \Sigma_\delta &= \Sigma_{P:G:s} + \Sigma_{PI:G:s} + \Sigma_{G:s} + \Sigma_{GI:s} + \Sigma_{sI} \\ &= \frac{1}{n_{p:g:s}}\Sigma_{p:g:s} + \frac{1}{n_{p:g:s}n_i}\Sigma_{pi:g:s} + \frac{1}{n_{g:s}}\Sigma_{g:s} + \frac{1}{n_{g:s}n_i}\Sigma_{gi:s} + \frac{1}{n_i}\Sigma_{si}. \end{aligned} \tag{7}$$

The absolute error matrix was obtained as

$$\begin{aligned}\Sigma_{\Delta} &= \Sigma_{P:G:s} + \Sigma_{PI:G:s} + \Sigma_{G:s} + \Sigma_{GI:s} + \Sigma_{sI} + \Sigma_{I} \\ &= \frac{1}{n_{p:g:s}}\Sigma_{p:g:s} + \frac{1}{n_{p:g:s}n_i}\Sigma_{pi:g:s} + \frac{1}{n_{g:s}}\Sigma_{g:s} + \frac{1}{n_{g:s}n_i}\Sigma_{gi:s} + \frac{1}{n_i}\Sigma_{si}. \\ &+ \frac{1}{n_i}\Sigma_{i}\end{aligned} \quad (8)$$

Error variances and coefficients are reported in Table 4. Relative error variance and absolute error variance were similar in magnitude for both pretest and posttest. This was expected, as the variance of items was very small for both administrations. Consequently, the estimates of $\hat{\Phi}$ and $\mathbf{E}\hat{\rho}^2$ were similar, as were the values of the signal-noise ratio using relative error ($S/N - Rel$) and using absolute error ($S/N - Abs$). Both $\hat{\rho}_{pre,post}(\delta)$ and $\hat{\rho}_{pre,post}(\Delta)$ were positive and moderate in magnitude.

As presented in Table 4, $\mathbf{E}\hat{\rho}^2_{diff}(s) = 0.364$, suggesting that the rank ordering of sites based on mean difference scores is likely to vary considerably. Similarly, the small value of $\hat{\Phi}_{diff}(s)$ suggests that for a randomly selected site, the absolute magnitude of the observed difference score is not a dependable estimate of the true difference score. S/N is the ratio of universe score variance to error variance and can range from 0 to infinity. Both $S/N - Rel$ and $S/N - Abs$ indicate that the universe score variance was half the error variance. This suggests that, under the current measurement conditions, detecting meaningful differences among sites is quite difficult.

2.2.2 $(p^{\bullet} : s^{\bullet})\,\mathbf{x}\,i^{\bullet}$ and $(P^{\bullet} : s^{\bullet})\,\mathbf{x}\,I^{\bullet}$

When the group facet was omitted, the G study design simplified to $(p^{\bullet} : s^{\bullet})\,\mathrm{x}\,i^{\bullet}$. Figure 2 provides a Venn diagram representation of this design. The number of persons within sites ($n_{p:s}$) ranged from 2 to 139, with a harmonic mean of 21.7. The number of items (n_i) was 25. Estimates of the G study variance and covariance components for both pretest and posttest scores are reported in Table 5.

For both pretest and posttest scores, the largest variance component, aside from the residual, was for $p : s$, which was expected, as it reflects the sum of two variance component $s : p$ and ps. $\hat{\sigma}^2_{pre}(s)$ and $\hat{\sigma}^2_{pre}(si)$ were similar in magnitude and represented the smallest among the pretest variance components. In addition, $\hat{\sigma}^2_{pre}(p : s)$ was almost 10 times larger than $\hat{\sigma}^2_{pre}(s)$, and $\hat{\sigma}^2_{post}(p : s)$ was almost 7 times larger than $\hat{\sigma}^2_{post}(s)$. This suggests that site-level coefficients may be smaller than person-level coefficients.

The variance components for s, $p : s$, si were all larger for posttest scores than for pretest scores. As shown in Table 5, $\hat{\sigma}^2_{post}(s)$ was nearly twice as large as $\hat{\sigma}^2_{pre}(s)$, whereas the increase in si variance from pretest ($\hat{\sigma}^2_{pre}(si) = 0.021$) to posttest ($\hat{\sigma}^2_{post}(si) = 0.026$) was smaller by comparison. Variance of si represent an interaction effect between s and i. Then both variance of s and variance of i have an impact on the variance of si. When $\hat{\sigma}^2_{post}(s)$ was notably larger

than $\hat{\sigma}^2_{pre}(s)$ while $\hat{\sigma}^2_{post}(i)$ was smaller than $\hat{\sigma}^2_{pre}(i)$, which is expected since the same items were used in both administrations, smaller increase in the variance of si compared to the increase in variance of s is reasonable. These trends were also found in the $(p^\bullet : g^\bullet : s^\bullet) \,\mathrm{x}\, i^\bullet$ design, reinforcing the conclusion that sites' posttest mean scores were considerably more variable than pretest mean scores, while rank order of item scores remained relatively consistent. The relative magnitudes of disattenuated correlations-$\hat{\rho}_{pre,post}(i)$, $\hat{\rho}_{pre,post}(si)$, and $\hat{\rho}_{pre,post}(s)$—as presented in Table 5, further support these findings.

The D study random effects design, with sites as the object of measurement, is denoted as $(P^\bullet : s^\bullet) \,\mathrm{x}\, I^\bullet$. The universe of generalization involves persons and items. The G study sample sizes were used as the corresponding D study sample sizes. The estimated D study variance and covariance components are summarized in Table 6. The universe score variance-covariance components for sites were $\hat{\sigma}^2_{pre}(\tau) = 0.024$ and $\hat{\sigma}^2_{post}(\tau) = 0.042$. The universe score correlation $\hat{\rho}_{pre,post}(\tau)$=0.543, as presented in Table 7, under this design was notably larger than that for the previous $(P^\bullet : G^\bullet : s^\bullet) \,\mathrm{x}\, i^\bullet$ design. This suggests that ignoring group facet resulted in inflated correlation estimates.

The D study variance and covariance components given in Table 6 were used to compute the error variances and coefficients, which are reported in Table 7. The relative and absolute error variances were similar in magnitude for both pretest and posttest scores. This was expected since $\hat{\sigma}^2(I)$ was very small for both pretest and posttest. When the group facet is ignored, the within-group variability contributes to the universe score variance, resulting in an increase in $\hat{\sigma}^2(\tau)$. Moreover, $\hat{\sigma}^2(\delta)$ and $\hat{\sigma}^2(\Delta)$ decreased in magnitude, which was expected since the variability of groups within sites contributes to the error variance. This resulted in increases in $\hat{\Phi}$, $\mathbf{E}\hat{\rho}^2$, $S/N - Rel$, and $S/N - Abs$.

$\mathbf{E}\hat{\rho}^2(x_{diff})$ was relatively high suggesting that the rank ordering of sites in terms of mean difference scores was consistent. Similarly, the relatively large value of $\hat{\Phi}(x_{diff})$ suggests that, for a randomly selected site, the absolute magnitude of observed difference score is a dependable estimate of the true difference score. Both $S/N - Rel$ and $S/N - Abs$ indicate that the universe score variance was 2.5 times larger than the error variance. This suggests that detecting differences among sites is relatively easier with this measurement procedure compared to $(P^\bullet : G^\bullet : s^\bullet) \,\mathrm{x}\, i^\bullet$. However, these coefficients were likely overestimated, as the data structure was misrepresented in this simplified design. When sites are the object of measurement, both persons and groups should be treated as facets in the universe of generalization. By omitting the group facet, this design failed to properly account for group-level variability, which contributes to the error variance, leading to inflated estimates of reliability coefficients.

2.2.3 $(p^\bullet : g^\bullet) \,\mathbf{x}\, i^\bullet$ and $(P^\bullet : g^\bullet) \,\mathbf{x}\, I^\bullet$

When the site facet was ignored, the G-study design became $(p^\bullet : g^\bullet) \,\mathrm{x}\, i^\bullet$ as illustrated in Figure 3. The number of persons within the groups ($n_{p:g}$) ranged from 2 to 86, with a harmonic mean of 8.119. The number of items (n_i) was 25. The estimated G study variance and covariance components for both pretest

and posttest scores are given in Table 8.

The variance components for g and $p:g$ were larger for posttest scores than for pretest scores, while i and $pi:g$ components were smaller for posttest scores. For both pretest and posttest scores, the largest variance component other than the residual was for $p:g$, which was expected since it reflects the combined variance of p and pg.

$\hat{\sigma}^2_{pre}(g)$ and $\hat{\sigma}^2_{pre}(gi)$ were similar in magnitude and the smallest among the pretest variances. $\hat{\sigma}^2_{pre}(p:g)$ was almost 7 times greater than $\hat{\sigma}^2_{pre}(g)$. The difference between pretest and posttest variances was larger for g than $p:g$. Specifically, $\hat{\sigma}^2_{post}(g)$ was larger than $\hat{\sigma}^2_{pre}(g)$, while $\hat{\sigma}^2_{post}(i)$ was substantially smaller than $\hat{\sigma}^2_{pre}(i)$. This suggests that group mean scores were more variable in the posttest, whereas item scores were more consistent in the posttest. The correlation between pretest and posttest items scores, $\hat{\rho}_{pre,post}(i)$, was very high, as observed in the previous designs. The correlations for g, $p:g$, and gi were moderate in magnitude.

The D study random effects design with groups as the object of measurement was $(P^\bullet : g^\bullet)\,\mathrm{x}\,I^\bullet$. Sites were ignored in this design. The estimated D study variance and covariance components are summarized in Table 9. The universe of generalization involved persons and items. In both the universe of generalization and D study design, items were crossed with groups, and persons were nested within groups.

Groups were the object of measurement in this design. The estimated universe score variance for group mean scores were $\hat{\sigma}^2_{pre}(g) = 0.037$ for the pretest and $\hat{\sigma}^2_{post}(g) = 0.062$ for the posttest. The universe score of a group represents the expected value of the group's mean score over persons and items. The correlation between group mean universe scores for pretest and posttest ($\hat{\rho}_{pre,post}(\tau) = 0.595$) was moderate in size and positive.

The D study variance and covariance components presented in Table 9 were used to compute the error variances and coefficients, which are reported in Table 10. Consistent with previous designs, the relative and absolute error variances were similar in magnitude for both pretest and posttest scores due to very small $\hat{\sigma}^2(I)$. This resulted in similar values between $\hat{\Phi}$ and $\mathbf{E}\hat{\rho}^2$, as well as between $S/N-Rel$ and $S/N-Abs$. Both coefficients, $\mathbf{E}\hat{\rho}^2(x_{diff}) = 0.582$ and $\hat{\Phi}(x_{diff}) = 0.5794$, were moderate in magnitude. Additionally, both $S/N-Rel$ and $S/N-Abs$ indicate that the universe score variance was quite low compared to the error variance. This suggests that detecting differences among groups is relatively difficult with this measurement procedure.

2.2.4 $g^\bullet\,\mathrm{x}\,i^\bullet$ and $g^\bullet\,\mathrm{x}\,I^\bullet$

When both the site and person facets were ignored, the G study design becomes $g^\bullet\,\mathrm{x}\,i^\bullet$ which is presented in Figure 4. The number of groups (n_g) was 89, and the number of items (n_i) was 25. Estimates of the G study variance and covariance components for both pretest and posttest scores are presented in Table 11. It was found that $\hat{\sigma}^2_{post}(g)$ was larger than $\hat{\sigma}^2_{pre}(g)$, while both $\hat{\sigma}^2_{pre}(i)$ and $\hat{\sigma}^2_{pre}(gi)$ were larger than $\hat{\sigma}^2_{post}(i)$ and $\hat{\sigma}^2_{post}(gi)$, respectively. $\hat{\rho}_{pre,post}(i)$ was very high

and $\hat{\rho}_{pre,post}(g)$ was moderate.

The D study random effects design, with groups as object of measurement, was $g^\bullet \times I^\bullet$. The universe of generalization involved items only. Estimates of the D study variance and covariance components are summarized in Table 12. The D study sample sizes were the same as those used in the G study. Groups were the object of measurement in this design. The universe score variances were $\hat{\sigma}^2_{pre}(\tau) = 0.064$ and $\hat{\sigma}^2_{post}(\tau) = 0.102$ for the groups' pretest and posttest scores, respectively. The correlation between group universe scores for the pretest and posttest was $\hat{\rho}_{pre,post}(\tau) = 0.505$.

The D study variance and covariance components presented in Table 12 were used to compute the error variances and coefficients, which are reported in Table 13. Similar to the previous designs, the relative and absolute error variances were similar to each other for both the pretest and posttest. Compared to the $(P^\bullet : g^\bullet) \times I^\bullet$ design, the values of $\mathbf{E}\hat{\rho}^2(x_{diff})$ and $\hat{\Phi}(x_{diff})$, $S/N - Rel$, and $S/N - Abs$ all increased. These inflated coefficients are likely attributable to the omission of the person facet in this design.

2.2.5 $p^\bullet \times i^\bullet$ and $p^\bullet \times I^\bullet$

The next design to consider is $p^\bullet \times i^\bullet$, in which both the site and group facets were omitted, and persons serve as the object of measurement. Figure 5 illustrates a Venn diagram representation. The number of persons (n_p) was 1517, and the number of items (n_i) was 25. Table 14 provides the estimated G study variance and covariance components. The patterns of variance components observed in this design were similar to those found in $g^\bullet \times i^\bullet$. Table 15 presents the variance and covariance components for the D study $p^\bullet \times I^\bullet$ design.

The universe score variances for persons were $\hat{\sigma}^2_{pre}(\tau) = 0.271$ and $\hat{\sigma}^2_{post}(\tau) = 0.326$. D study results are reported in Table 16. Consistent with the $g^\bullet \times i^\bullet$ design, $\hat{\sigma}^2_{post}(\delta)$ was smaller than $\hat{\sigma}^2_{pre}(\delta)$. A similar pattern was observed for absolute error variances, with $\hat{\sigma}^2_{post}(\Delta)$ being smaller than $\hat{\sigma}^2_{pre}(\Delta)$. All coefficients, $\mathbf{E}\hat{\rho}^2(x_{diff})$, $\hat{\Phi}(x_{diff})$, $S/N - Rel$, $S/N - Abs$, were relatively large, indicating a high level of consistency and precision.

2.2.6 $s^\bullet \times i^\bullet$ and $s^\bullet \times I^\bullet$

In the G study design $s^\bullet \times i^\bullet$, both the group and person facets were omitted, and sites served as the object of measurement (see Figure 6). The number of sites (n_s) was 32, and the number of items (n_i) was 25. The G study variance and covariance components are given in Table 17. Note that $\hat{\sigma}^2_{post}(s) > \hat{\sigma}^2_{pre}(s)$, $\hat{\sigma}^2_{pre}(i) > \hat{\sigma}^2_{post}(i)$, and $\hat{\sigma}^2_{pre}(si) > \hat{\sigma}^2_{post}(si)$. In addition, $\hat{\rho}_{pre,post}(i)$ is very large, while $\hat{\rho}_{pre,post}(s)$ is moderate in magnitude. The variance and covariance components for the D study $s^\bullet \times I^\bullet$ are illustrated in Table 18.

The correlation of site universe scores between pretest and posttest was $\hat{\rho}_{pre,post}(\tau) = 0.579$. The error variances and coefficients are reported in Table 19. Consistent with the $g^\bullet \times i^\bullet$ and $p^\bullet \times i^\bullet$ designs, $\hat{\sigma}^2_{post}(\delta)$ was smaller than $\hat{\sigma}^2_{pre}(\delta)$. Similarly, the same pattern was observed between $\hat{\sigma}^2_{pre}(\Delta)$ and

$\hat{\sigma}^2_{post}(\Delta)$. Relatively large values were obtained for all coefficients: $\mathbf{E}\hat{\rho}^2(x_{diff})$ and $\hat{\Phi}(x_{diff})$, $S/N - Rel$ and $S/N - Abs$. Again, the person and group facets were not explicitly modeled in this design, which may have led to inflated estimates of these coefficients.

2.2.7 Comparison across Designs

There are three designs in which sites serve as the object of measurement: $(P^\bullet : G^\bullet : s^\bullet) \times I^\bullet$, $(P^\bullet : s^\bullet) \times I^\bullet$, and $s^\bullet \times I^\bullet$. The corresponding coefficients for these designs are summarized in Table 20. A clear trend was observed in both the relative error variances ($\hat{\sigma}^2(\delta)$) and the absolute error variances ($\hat{\sigma}^2(\Delta)$), with the smallest value occurring in the $s^\bullet \times I^\bullet$ design and the largest in the $(P^\bullet : G^\bullet : s^\bullet) \times I^\bullet$ design. The increase in the $\hat{\sigma}^2(\delta)$ and $\hat{\sigma}^2(\Delta)$ accompanied by a decrease in $\hat{\sigma}^2(\tau)$ resulted in a decrease in the generalizability coefficient (i.e., $\mathbf{E}\hat{\rho}^2$), the dependability coefficient (i.e., $\hat{\Phi}$), $S/N - Rel$, and $S/N - Abs$. These patterns illustrate how omission of person and group facets can influence the estimation of reliability-related indices.

As the relative error variance increased across $s^\bullet \times I^\bullet$, $(P^\bullet : s^\bullet) \times I^\bullet$, and $(P^\bullet : G^\bullet : s^\bullet) \times I^\bullet$, so did the relative error correlations. As the absolute error variance increased, the absolute error correlations decreased. As summarized in Table 1, there are more variance components contributed to the relative and absolute error variances across these designs. This resulted in an increase in both error variances. To be able to understand this trend in error correlations, we need to look at how they were computed.

In all of these designs, the computational definition of relative error variance and covariance differs from the absolute error variance by only the I facet. For instance, in the $s^\bullet \times I^\bullet$ design, $\hat{\rho}_{pre,post}(\delta)$ and $\hat{\rho}_{pre,post}(\Delta)$ computations differed by addition of $\hat{\sigma}_{pre,post}(I)$ in the numerator and addition of $\hat{\sigma}^2(I)$ to the both of the standard deviations in the denominator. As can be seen in Equations 9 and 10, in the denominator, addition of $\hat{\sigma}^2(I)$ to $\hat{\sigma}^2(\delta)$ equaled $\hat{\sigma}^2(\Delta)$, and then absolute error variances were multiplied. Furthermore, since the ignored facets were within the object of measurement, the variance and covariance components associated with the I facet, $\hat{\sigma}^2_{post}(I)$, $\hat{\sigma}^2_{pre}(I)$, and $\hat{\sigma}_{pre,post}(I)$, remained the same. These identical values were added to the terms used in computing the correlations. Because the other variance components differed across these designs, the addition of the same values seemed to have a greater impact on the correlation value. Specifically, the product of the standard deviations may have increased more than the covariance, thereby reducing the correlation values. This led to the decrease in $\hat{\rho}_{pre,post}(\Delta)$ across designs:

$$\hat{\rho}_{pre,post}(\delta) = \frac{\hat{\sigma}_{pre,post}(\delta)}{\sqrt{\left(\hat{\sigma}^2_{pre}(\delta)\right)\left(\hat{\sigma}^2_{post}(\delta)\right)}} = \frac{\hat{\sigma}_{pre,post}(sI)}{\sqrt{\left(\hat{\sigma}^2_{pre}(sI)\right)\left(\hat{\sigma}^2_{post}(sI)\right)}}. \quad (9)$$

$$\begin{aligned}\hat{\rho}_{pre,post}(\Delta) &= \frac{\hat{\sigma}_{pre,post}(\Delta)}{\sqrt{\left(\hat{\sigma}^2_{pre}(\Delta)\right)\left(\hat{\sigma}^2_{post}(\Delta)\right)}} \\ &= \frac{\hat{\sigma}_{pre,post}(sI) + \hat{\sigma}_{pre,post}(I)}{\sqrt{\left(\hat{\sigma}^2_{pre}(sI) + \hat{\sigma}^2_{pre}(I)\right)\left(\hat{\sigma}^2_{post}(sI) + \hat{\sigma}^2_{post}(I)\right)}} \\ &= \frac{\hat{\sigma}_{pre,post}(\delta) + \hat{\sigma}_{pre,post}(I)}{\sqrt{\left(\hat{\sigma}^2_{pre}(\delta) + \hat{\sigma}^2_{pre}(I)\right)\left(\hat{\sigma}^2_{post}(\delta) + \hat{\sigma}^2_{post}(I)\right)}}. \end{aligned} \tag{10}$$

In the $(P^\bullet : G^\bullet : s^\bullet)\,\mathrm{x}\,I^\bullet$ design, the nesting indices of persons and groups were explicitly treated as facets in the universe of generalization, allowing their contributions to the error variance to be separately estimated. Similarly, in the $(P^\bullet : s^\bullet)\,\mathrm{x}\,I^\bullet$ design, persons were modeled as a facet, and their variability was explicitly included in the estimation of error variance. In contrast, the $s^\bullet\,\mathrm{x}\,I^\bullet$ design did not specify any nesting indices, meaning that variability due to persons and groups was absorbed into the site component. Although these sources of variability still contribute to the total error variance, their impact is not directly estimated and may be underrepresented or misallocated. These differences in design specification help explain the increase in error variances, change in error correlations, and decrease in generalizability and dependability coefficients as the model becomes more simplified and omits key facets of the data structure.

Table 21 shows how the coefficients change when the object of measurement changes across persons, groups, and sites. Relative and absolute error correlations increased, universe score variances decreased, and reliability-related coefficients decreased across $p^\bullet\,\mathrm{x}\,I^\bullet$, $(P^\bullet : g^\bullet)\,\mathrm{x}\,I^\bullet$, $(P^\bullet : G^\bullet : s^\bullet)\,\mathrm{x}\,I^\bullet$, respectively. In terms of $\hat{\sigma}^2(\delta)$ and $\hat{\sigma}^2(\Delta)$, there was not a clear pattern across designs. This is reasonable because D study variance components and their divisors varied depending on the specified model.

3 Discussion

This study examines the reliability of difference scores when they are obtained from multilevel pretest-posttest data. The nature of difference scores (i.e., composite scores), combined with the multilevel data structure, necessitates considering both variance and covariance components in reliability estimation. Generalizability theory, with its univariate, multivariate, and extended multivariate methodologies, provides a flexible and powerful framework for modeling reliability in such complex contexts (Brennan, 2001a; Brennan et al., 2022). In this study, the following G study designs were investigated: $(p^\bullet : g^\bullet : s^\bullet)\,\mathrm{x}\,i^\bullet$, $(p^\bullet : s^\bullet)\,\mathrm{x}\,i^\bullet$, $(p^\bullet : g^\bullet)\,\mathrm{x}\,i^\bullet$, $g^\bullet\,\mathrm{x}\,i^\bullet$, $p^\bullet\,\mathrm{x}\,i^\bullet$, and $s^\bullet\,\mathrm{x}\,i^\bullet$. A central issue addressed in this study is how values of reliability coefficients, representing the magnitude and the rank order stability of difference scores, vary across different objects of

measurement and across designs with neglected nested facets by decomposing reliability of difference scores into its constituting facets.

When the focus is on decision-making regarding relative standing of the unit of analysis, the generalizability coefficient $\mathbf{E}\hat{\rho}^2$ is more appropriate to use (Brennan & Kane, 1977; Miller & Kane, 2001). In contrast, when the goal is to make a criterion-based decision, the dependability coefficient $\hat{\Phi}$ is more suitable. In this study, the values of $\mathbf{E}\hat{\rho}^2$ and $\hat{\Phi}$ were consistently close to each other across all designs, and this pattern held for pretest, posttest, and difference scores. In the designs used, the definition of absolute error variance differs from relative error variance only by the inclusion of the item facet. Consistent across all designs, the variance associated with the item facet was relatively small. Furthermore, the estimated disattenuated correlation between pretest and posttest scores for items was consistently high in all the designs. These findings help explain why $\mathbf{E}\hat{\rho}^2$ and $\hat{\Phi}$ coefficients were similar in magnitude.

Both $\mathbf{E}\hat{\rho}^2$ and $\hat{\Phi}$ coefficients were largest when the object of measurement was persons (i.e., students) and smallest when the object of measurement was sites. One reason behind this finding was the low sample sizes associated with groups within sites (i.e., $g : s$) and persons within groups within sites (i.e., $p : g : s$) facets. Total number of persons was 1517 whereas the number of persons per group varied across groups within sites with a harmonic mean of 8.277, and the number of groups per site varied across sites with a harmonic mean of 2.349. If the number of persons per group per site or the number of groups per site increases, then $\mathbf{E}\hat{\rho}^2$ and $\hat{\Phi}$ coefficients of sites will increase since these numbers were divisors in the computation of $\mathbf{E}\hat{\rho}^2$ and $\hat{\Phi}$ coefficients. Furthermore, $\mathbf{E}\hat{\rho}^2$ values for pretest scores were consistently smaller than those for posttest scores, a pattern that also held for $\hat{\Phi}$. These findings are consistent with those reported by Miller and Kane (2001). One possible explanation for these patterns is the meaningful change observed between pretest and posttest scores, which may have led to increased universe score variance, and, in turn, higher reliability estimates in the posttest condition.

When the object of measurement is sites, both groups and persons should be treated as facets in the universe of generalization. Sites often differ in terms of culture, background, and learning conditions. Moreover, in this study, the number of groups within sites and the number of persons within groups varied substantially. Moreover, when these facets are not explicitly included in the estimation, there is only site means left that can be used for reliability estimation. That means the variance of the persons within groups within sites and of the groups within sites could not be completely included in the reliability estimation. In other words, the information provided by the persons and groups are lost to some extent resulting in inflated reliability coefficients. This underscores the importance of incorporating the nested structure of the sampling design into reliability estimation. Specifically, when making decisions at the group level, person-level variance should not be ignored. Similarly, when making decisions at the site level, group-level variance must be accounted for to avoid misleading reliability estimates.

This study decomposed the reliability of site mean difference scores and the

associated error correlations into their constituent facets via multivariate G theory. By comparing different designs with the same object of measurement, we observed that the pattern of correlated errors varied across designs. Specifically, across the $s^{\bullet} \times I^{\bullet}$, $(P^{\bullet} : s^{\bullet}) \times I^{\bullet}$, and $(P^{\bullet} : G^{\bullet} : s^{\bullet}) \times I^{\bullet}$ designs, the relative error correlations increased, whereas the absolute error correlations decreased. This finding can be explained by the fact that the absolute error includes more facets than the relative error by definition. Consequently, absolute error correlations may decrease when the product of the absolute error standard deviations increases faster than the absolute error covariance. This study provided empirical evidence for this finding.

4 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. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement*, 32(4), 385-396. https://doi.org/10.1111/j.1745-3984.1995.tb00473.x

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

Brennan, R. L. (2001b). *Manual for mGENOVA*. [Computer software and manual]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on https://education.uiowa.edu/casma/computer-programs).

Brennan, R. L. (2001c). *Manual for urGENOVA*. [Computer software and manual]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on https://education.uiowa.edu/casma/computer-programs).

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, 14(3), 277–289. http://www.jstor.org/stable/1434319

Brennan, R. L., Kim, S. Y., & Lee, W. (2022). Extended multivariate generalizability theory with complex design structures. *Educational and Psychological Measurement*, 82(4), 617–642. https://doi.org/10.1177/001316442110497

Brennan, R. L., Yin, P., & Kane, M. T. (2003). Methodology for examining the reliability of group mean difference scores. *Journal of Educational Measurement*, 40, 207–230. https://doi.org/10.1111/j.1745-3984.2003.tb01105.x

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, 57(3), 373–399. https://doi.org/10.1177/0013164497057003001

Gu, Z., Emons, W. H. M. & Sijtsma, K. (2018). Review of issues about classical change scores: A multilevel modeling perspective on some enduring beliefs. *Psychometrika*, 83, 674–695. https://doi.org/10.1007/s11336-018-9611-3

Kane, M. (1996). The Precision of measurements. *Applied Measurement in Education*, 9(4), 355–379. https://doi.org/10.1207/s15324818ame0904_4

Miller, T. B., & Kane, M. (2001). The precision of change scores under absolute and relative interpretations. *Applied Measurement in Education*, 14(4), 307-327. https://doi.org/10.1207/S15324818AME1404_1

Wade, R. & Yarbrough, D. (2007). Service–learning in the social studies: Civic outcomes of the 3rd–12th grade CiviConnections program. *Theory & Research in Social Education*, 35:3, 366-392. http://dx.doi.org/10.1080/00933104.2007.10473341

Wei, X., & Haertel, E. (2011). The effect of ignoring classroom-level variance in estimating the generalizability of school mean scores. *Educational Measurement: Issues and Practice*, 30(1), 13-22. https://doi.org/10.1111/j.1745-3992.2010.00196.x

Yin, P., & Brennan, R. L. (2002). An investigation of difference scores for a grade-level testing program. *International Journal of Testing*, 2(2), 83. https://doi.org/10.1207/S15327574IJT0202_1

Table 1: Variance components across designs

	$\sigma^2(\tau)$	$\sigma^2(\delta)$	$\sigma^2(\Delta)$
$(p^\bullet : g^\bullet : s^\bullet) \times i^\bullet$	s	$p:g:s,\ pi:g:s,\ g:s,\ gi:s,\ si$	$p:g:s,\ pi:g:s,\ g:s,\ gi:s,\ si,\ i$
$(p^\bullet : s^\bullet) \times i^\bullet$	s	$p:s,\ pi:s,\ si$	$p:s,\ pi:s,\ si,\ i$
$(p^\bullet : g^\bullet) \times i^\bullet$	g	$p:g,\ pi:g,\ gi$	$p:g,\ pi:g,\ gi,\ i$
$g^\bullet \times i^\bullet$	g	gi	$gi,\ i$
$p^\bullet \times i^\bullet$	p	pi	$pi,\ i$
$s^\bullet \times i^\bullet$	s	si	$si,\ i$

Table 2: G study variance and covariance components for $(p^\bullet : g^\bullet : s^\bullet) \times i^\bullet$

α	$\hat{\sigma}^2_{pre}$	$\hat{\sigma}^2_{post}$	$\hat{\sigma}_{pre,post}$	$\hat{\rho}_{pre,post}$
s	0.013	0.024	0.008	0,432
$g:s$	0.025	0.039	0.021	0,687
$p:g:s$	0.234	0.265	0.147	0,589
i	0.130	0.082	0.102	0,983
si	0.011	0.019	0.011	0,766
$gi:s$	0.023	0.016	0.009	0,479
$pi:g:s$	0.577	0.491	0.133	0,250

Table 3: D study variance and covariance components for $(P^\bullet : G^\bullet : s^\bullet) \times I^\bullet$

n	α	$\hat{\sigma}^2_{pre}$	$\hat{\sigma}^2_{post}$	$\hat{\sigma}_{pre,post}$
	s	0.013	0.024	0.008
2.349	$G:s$	0.011	0.016	0.009
8.277	$P:G:s$	0.028	0.032	0.018
25.0	I	0.005	0.003	0.004
25.0	sI	0.000	0.001	0.000
58.716	$GI:s$	0.000	0.000	0.000
206.924	$PI:G:s$	0.003	0.002	0.001

Table 4: D study results for $(P^\bullet : G^\bullet : s^\bullet) \times I^\bullet$

	x_{pre}	x_{post}	x_{diff}	$\hat{\rho}_{pre,post}$
$\hat{\sigma}^2(\tau)$	0.013	0.024	0.022	0.432
$\hat{\sigma}^2(\delta)$	0.043	0.052	0.038	0.597
$\hat{\sigma}^2(\Delta)$	0.048	0.055	0.039	0.626
$\mathbf{E}\hat{\rho}^2$	0.232	0.319	0.364	
$\hat{\Phi}$	0.212	0.306	0.362	
S/N-Rel	0.303	0.468	0.572	
S/N-Abs	0.270	0.440	0.566	

Table 5: G study variance and covariance components for $(p^\bullet : s^\bullet) \times i^\bullet$

α	$\hat{\sigma}^2_{pre}$	$\hat{\sigma}^2_{post}$	$\hat{\sigma}_{pre,post}$	$\hat{\rho}_{pre,post}$
s	0.024	0.042	0.017	0,543
$p : s$	0.248	0.287	0.159	0,595
i	0.130	0.082	0.102	0,983
si	0.021	0.026	0.015	0,642
$pi : s$	0.589	0.500	0.138	0,255

Table 6: D study variance and covariance components for $(P^\bullet : s^\bullet) \times I^\bullet$

n	α	$\hat{\sigma}^2_{pre}$	$\hat{\sigma}^2_{post}$	$\hat{\sigma}_{pre,post}$
	s	0.024	0.042	0.017
21.73439	$P : s$	0.011	0.013	0.007
25.0	I	0.005	0.003	0.004
25.0	sI	0.001	0.001	0.001
543.3598	$PI : s$	0.001	0.001	0.000

Table 7: D study results for $(P^\bullet : s^\bullet) \times I^\bullet$

	x_{pre}	x_{post}	x_{diff}	$\hat{\rho}_{pre,post}$
$\hat{\sigma}^2(\tau)$	0.024	0.042	0.031	0.543
$\hat{\sigma}^2(\delta)$	0.013	0.015	0.012	0.574
$\hat{\sigma}^2(\Delta)$	0.019	0.018	0.013	0.661
$\mathbf{E}\hat{\rho}^2$	0.643	0.733	0.720	
$\hat{\Phi}$	0.564	0.693	0.714	
S/N-Rel	1.800	2.749	2.569	
S/N-Abs	1.296	2.258	2.495	

Table 8: G study variance and covariance components for $(p^\bullet : g^\bullet) \times i^\bullet$

α	$\hat{\sigma}^2_{pre}$	$\hat{\sigma}^2_{post}$	$\hat{\sigma}_{pre,post}$	$\hat{\rho}_{pre,post}$
g	0.037	0.062	0.029	0,595
$p : g$	0.234	0.265	0.147	0,589
i	0.130	0.083	0.102	0,982
gi	0.034	0.034	0.020	0,584
$pi : g$	0.577	0.491	0.133	0,250

Table 9: D study variance and covariance components for $(P^\bullet : g^\bullet) \times I^\bullet$

n	α	$\hat{\sigma}^2_{pre}$	$\hat{\sigma}^2_{post}$	$\hat{\sigma}_{pre,post}$
	g	0.037	0.062	0.029
8.119	$P : g$	0.029	0.033	0.018
25.0	I	0.005	0.003	0.004
25.0	gI	0.001	0.001	0.001
202.972	$PI : g$	0.003	0.002	0.001

Table 10: D study results for $(P^{\bullet} : g^{\bullet})\,\mathrm{x}\,I^{\bullet}$

	x_{pre}	x_{post}	x_{diff}	$\hat{\rho}_{pre,post}$
$\hat{\sigma}^2(\tau)$	0.037	0.062	0.042	0.595
$\hat{\sigma}^2(\delta)$	0.033	0.036	0.030	0.563
$\hat{\sigma}^2(\Delta)$	0.038	0.040	0.031	0.605
$\mathbf{E}\hat{\rho}^2$	0.531	0.631	0.582	
$\hat{\Phi}$	0.495	0.611	0.579	
S/N-Rel	1.132	1.711	1.390	
S/N-Abs	0.978	1.568	1.373	

Table 11: G study variance and covariance components for $g^{\bullet}\,\mathrm{x}\,i^{\bullet}$

α	$\hat{\sigma}^2_{pre}$	$\hat{\sigma}^2_{post}$	$\hat{\sigma}_{pre,post}$	$\hat{\rho}_{pre,post}$
g	0.064	0.102	0.041	0,505
i	0.135	0.085	0.105	0,979
gi	0.129	0.107	0.046	0,389

Table 12: D study variance and covariance components for $g^{\bullet}\,\mathrm{x}\,I^{\bullet}$

n	α	$\hat{\sigma}^2_{pre}$	$\hat{\sigma}^2_{post}$	$\hat{\sigma}_{pre,post}$
	g	0.064	0.102	0.041
25.0	I	0.005	0.003	0.004
25.0	gI	0.005	0.004	0.002

Table 13: D study results for $g^{\bullet}\,\mathrm{x}\,I^{\bullet}$

	x_{pre}	x_{post}	x_{diff}	$\hat{\rho}_{pre,post}$
$\hat{\sigma}^2(\tau)$	0.064	0.102	0.084	0.505
$\hat{\sigma}^2(\delta)$	0.005	0.004	0.006	0.389
$\hat{\sigma}^2(\Delta)$	0.011	0.008	0.006	0.669
$\mathbf{E}\hat{\rho}^2$	0.925	0.960	0.936	
$\hat{\Phi}$	0.858	0.930	0.931	
S/N-Rel	12.350	23.760	14.564	
S/N-Abs	6.026	13.286	13.594	

Table 14: G study variance and covariance components for $p^\bullet \times i^\bullet$

α	$\hat{\sigma}^2_{pre}$	$\hat{\sigma}^2_{post}$	$\hat{\sigma}_{pre,post}$	$\hat{\rho}_{pre,post}$
p	0.271	0.326	0.175	0,589
i	0.131	0.083	0.102	0,979
pi	0.610	0.525	0.153	0,270

Table 15: D study variance and covariance components for $p^\bullet \times I^\bullet$

n	α	$\hat{\sigma}^2_{pre}$	$\hat{\sigma}^2_{post}$	$\hat{\sigma}_{pre,post}$
	p	0.271	0.326	0.175
25.0	I	0.005	0.003	0.004
25.0	pI	0.024	0.021	0.006

Table 16: D study results for $p^\bullet \times I^\bullet$

	x_{pre}	x_{post}	x_{diff}	$\hat{\rho}_{pre,post}$
$\hat{\sigma}^2(\tau)$	0.271	0.326	0.247	0.589
$\hat{\sigma}^2(\delta)$	0.024	0.021	0.033	0.270
$\hat{\sigma}^2(\Delta)$	0.030	0.024	0.034	0.380
$\mathbf{E}\hat{\rho}^2$	0.917	0.940	0.882	
$\hat{\Phi}$	0.901	0.931	0.881	
S/N-Rel	11.113	15.551	7.458	
S/N-Abs	9.151	13.417	7.372	

Table 17: G study variance and covariance components for $s^\bullet \times i^\bullet$

α	$\hat{\sigma}^2_{pre}$	$\hat{\sigma}^2_{post}$	$\hat{\sigma}_{pre,post}$	$\hat{\rho}_{pre,post}$
s	0.051	0.079	0.037	0.579
i	0.126	0.077	0.096	0.977
si	0.073	0.054	0.026	0.412

Table 18: D study variance and covariance components for $s^\bullet \times I^\bullet$

n	α	$\hat{\sigma}^2_{pre}$	$\hat{\sigma}^2_{post}$	$\hat{\sigma}_{pre,post}$
	s	0.051	0.079	0.037
25.0	I	0.005	0.003	0.004
25.0	sI	0.003	0.002	0.001

Table 19: D study results for $s^\bullet \times I^\bullet$

	x_{pre}	x_{post}	x_{diff}	$\hat{\rho}_{pre,post}$
$\hat{\sigma}^2(\tau)$	0.051	0.079	0.057	0.57
$\hat{\sigma}^2(\delta)$	0.003	0.002	0.003	0.412
$\hat{\sigma}^2(\Delta)$	0.008	0.005	0.003	0.757
$\mathbf{E}\hat{\rho}^2$	0.946	0.974	0.950	
$\hat{\Phi}$	0.866	0.938	0.943	
S/N-Rel	17.564	36.966	18.851	
S/N-Abs	6.446	15.164	16.543	

Table 20: Summary of D study results for site mean difference scores across designs

	$s^\bullet \times I^\bullet$		$(P^\bullet : s^\bullet) \times I^\bullet$		$(P^\bullet : G^\bullet : s^\bullet) \times I^\bullet$	
	x_{diff}	$\hat{\rho}_{pre,post}$	x_{diff}	$\hat{\rho}_{pre,post}$	x_{diff}	$\hat{\rho}_{pre,post}$
$\hat{\sigma}^2(\tau)$	0.057	0.579	0.031	0.543	0.022	0.432
$\hat{\sigma}^2(\delta)$	0.003	0.412	0.012	0.574	0.038	0.597
$\hat{\sigma}^2(\Delta)$	0.003	0.757	0.013	0.661	0.039	0.626
$\mathbf{E}\hat{\rho}^2$	0.950		0.720		0.364	
$\hat{\Phi}$	0.943		0.714		0.362	
S/N-Rel	18.851		2.569		0.572	
S/N-Abs	16.543		2.495		0.566	

Table 21: Summary of D study results across different objects of measurement

	$p^\bullet \times I^\bullet$		$(P^\bullet : g^\bullet) \times I^\bullet$		$(P^\bullet : G^\bullet : s^\bullet) \times I^\bullet$	
	x_{diff}	$\hat{\rho}_{pre,post}$	x_{diff}	$\hat{\rho}_{pre,post}$	x_{diff}	$\hat{\rho}_{pre,post}$
$\hat{\sigma}^2(\tau)$	0.247	0.589	0.042	0.595	0.022	0.432
$\hat{\sigma}^2(\delta)$	0.033	0.270	0.030	0.563	0.038	0.597
$\hat{\sigma}^2(\Delta)$	0.034	0.380	0.031	0.605	0.039	0.626
$\mathbf{E}\hat{\rho}^2$	0.882		0.582		0.364	
$\hat{\Phi}$	0.881		0.579		0.362	
S/N-Rel	7.458		1.390		0.572	
S/N-Abs	7.372		1.373		0.566	

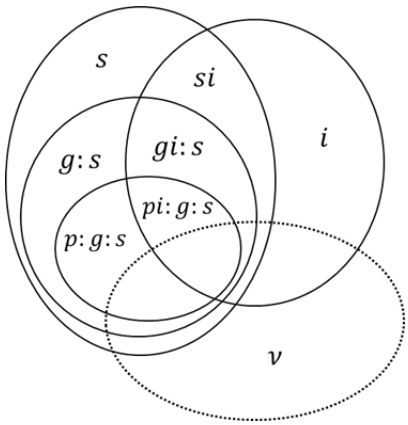

Figure 1: Venn diagram of $(p^{\bullet} : g^{\bullet} : s^{\bullet}) \times i^{\bullet}$

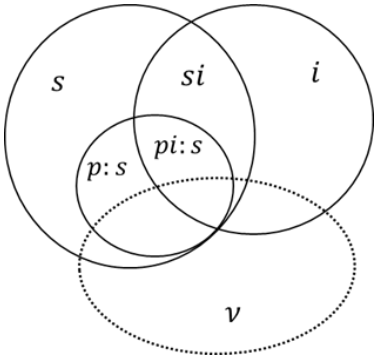

Figure 2: Venn diagram of $(p^{\bullet} : s^{\bullet}) \times i^{\bullet}$

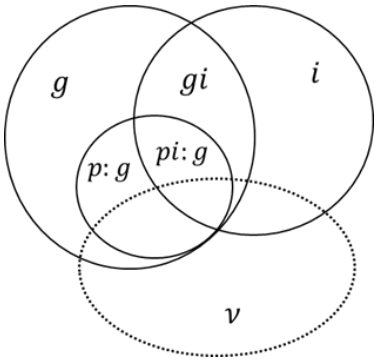

Figure 3: Venn diagram of $(p^{\bullet} : g^{\bullet}) \times i^{\bullet}$

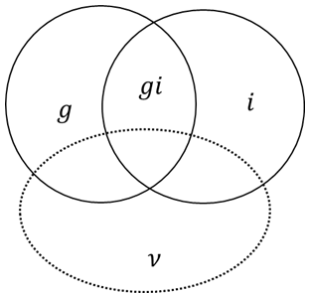

Figure 4: Venn diagram of $g^{\bullet} \times i^{\bullet}$

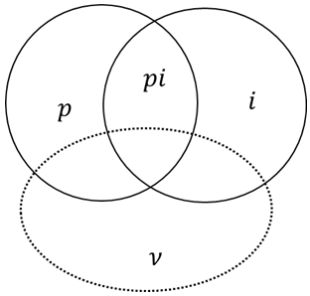

Figure 5: Venn diagram of $p^{\bullet} \times i^{\bullet}$

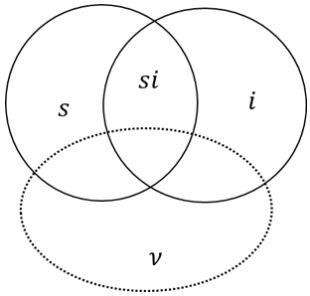

Figure 6: Venn diagram of $s^{\bullet} \times i^{\bullet}$