Analyzing complete generalizability theory designs using structural equation models

Vispoel, W.P.; Hong, H.; Lee, H.; Jorgensen, T.D.

DOI
10.1080/08957347.2023.2274573

Publication date
2023

Document Version
Final published version

Published in
Applied Measurement in Education

APPLIED MEASUREMENT IN EDUCATION
2023, VOL. 36, NO. 4, 372–393
https://doi.org/10.1080/08957347.2023.2274573

Routledge
Taylor & Francis Group

Analyzing Complete Generalizability Theory Designs Using Structural Equation Models

Walter P. Vispoel [a], Hyeri Hong [b], Hyeryung Lee [a], and Terrence D. Jorgensen [c]

[a]Department of Psychological and Quantitative Foundations, University of Iowa; [b]Department of Curriculum and Instruction, California State University Fresno; [c]Research Institute for Child Development and Education, University of Amsterdam

ABSTRACT

We illustrate how to analyze complete generalizability theory (GT) designs using structural equation modeling software (*lavaan* in R), compare results to those obtained from numerous ANOVA-based packages, and apply those results in practical ways using data obtained from a large sample of respondents, who completed the Self-Perception Profile for College Students (Neemann & Harter, 2012) on two occasions. Results revealed that estimates of variance components, generalizability coefficients, dependability coefficients, and proportions of measurement error derived from *lavaan* were essentially equivalent to those produced by the GT packages GENOVA and *gtheory* in R and variance component programs in SPSS, SAS, and R. Within the article and extended online Supplemental Material, we illustrate how indices obtained from these resources can be used for either norm- and criterion-referencing purposes and for estimating effects of changes made to measurement procedures. We further describe ways to use structural equation models for applications of GT beyond what conventional ANOVA-based packages would typically permit.

Generalizability theory (GT; Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963; Shavelson & Webb, 1991) represents a comprehensive framework for evaluating score consistency for both norm and criterion referencing purposes that allows for estimation of multiple sources of measurement error. Although variance components for GT designs have traditionally been derived using analysis of variance (ANOVA) procedures, they also can be obtained from structural equation models (SEMs; see, e.g., Ark, 2015; Jorgensen, 2021; Marcoulides, 1996; Morris, 2020; Raykov & Marcoulides, 2006; Vispoel, Hong, & Lee, 2023; Vispoel, Lee, Chen, & Hong, 2023a, 2023b, 2023c; Vispoel, Lee, & Hong, 2023; Vispoel, Lee, Xu, & Hong, 2022, 2023; Vispoel, Morris, & Kilinc, 2018a, 2018b, 2019; Vispoel, Xu, & Kilinc, 2021; Vispoel, Xu, & Schneider, 2022a). In this article, we use data obtained from a large sample of respondents who completed the Self-Perception Profile for College Students (Neemann & Harter, 2012) to illustrate how complete GT designs can be analyzed using structural equation models (SEMs), how results from such models compare to ones obtained using a variety of ANOVA-based procedures, and how those results can be applied effectively in practice.

CONTACT Walter P. Vispoel walter-vispoel@uiowa.edu Department of Psychological and Quantitative Foundations, 361 Lindquist Center, University of Iowa, Iowa City, IA, USA, 52242.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/08957347.2023.2274573

2. Background

2.1. GT Basics

When conducting GT analyses, researchers often distinguish between *generalizability* and *decision* studies. Within generalizability studies, *objects of measurement* and *universes of admissible observations* are identified, data are collected, and relevant *variance components* are estimated. In our examples of GT analyses, *persons* are the objects of measurement, and *items* and/or *occasions* serve as possible universes of generalization. Each universe of generalization is represented as a *facet* corresponding to a source of measurement error within the GT design. Explained systematic variance unrelated to measurement error within GT designs is called *universe* or *person score variance* and conceptually parallels *true score variance* in classical test theory and *communality* in factor analysis. Universe scores for the objects of measurement would represent estimates of average results obtained over all possible independent replications of the assessment procedure within the domains of interest. Corresponding errors of measurement within the GT design, in turn, would reflect limitations in generalizing results to those broader domains.

In decision studies, variance components derived in generalizability studies are used to estimate indices of reliability for norm- and/or criterion-referencing purposes based on the original generalizability study design or ones altered within the decision study for possible changes made to a measurement procedure. Common applications within decision studies include estimating relevant indices of generalizability, dependability, and measurement error and how those indices might change when altering numbers of conditions for facets and/or restricting the original universes of generalization to a lesser number of facets (see, e.g., Brennan, 2001; Vispoel, 2023a; Vispoel, Hong, & Lee, 2023; Vispoel, Lee, Chen, & Hong, 2023a, 2023b, 2023c; Vispoel, Morris, & Kilinc, 2018a, 2018b, 2018c, 2018d, 2019; Vispoel & Tao, 2013; Vispoel, Xu, & Schneider, 2022a, 2022b). Although we discuss common GT designs using items and occasions from objectively scored measures as universes of generalization here, the same techniques can be readily applied to subjectively scored measures (e.g., performance assessments) and other facets (raters, alternative tasks, test forms, etc.).

2.2. Single Facet Designs

We begin with a *persons by items* ($p \times i$) random-facet design, and show in Equations 1 and 2 how estimated observed score variance is partitioned for individual items and item means (i.e., composites forming by averaging item scores). The letter I is capitalized in Equation 2 and elsewhere to signify such averaging over item scores. The prime over n_i in the equation indicates that numbers of items can be further altered for possible changes made to the measurement procedure in a decision study.

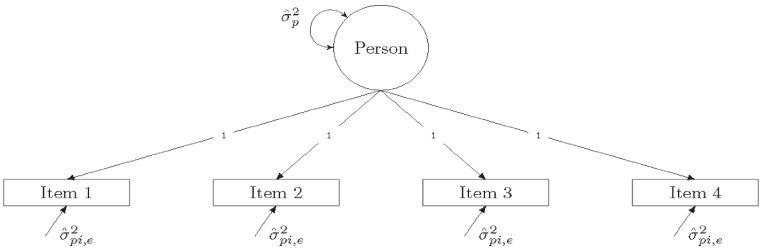
$$p \times i \text{ design Individual item score level}: \hat{\sigma}^2_{Y_{pi}} = \hat{\sigma}^2_p + \hat{\sigma}^2_{pi,e} + \hat{\sigma}^2_i, \tag{1}$$

$$p \times I \text{ design Mean item score level}: \hat{\sigma}^2_{Y_{pI}} = \hat{\sigma}^2_p + \frac{\hat{\sigma}^2_{pi,e}}{n'_i}, \tag{2}$$

where $\hat{\sigma}^2$ = estimated variance component, Y_{pi} = score for a particular person on a given item, Y_{pI} = mean across all items for a particular person, and n'_i = number of items.

In traditional applications of GT, the variance components represented in Equations 1 and 2 would be estimated using ANOVA-based procedures. These estimates can be obtained from customized packages designed specifically for GT analyses (e.g., GENOVA, Crick & Brennan, 1983; *gtheory* package in R, Moore, 2016) or from variance component programs within comprehensive statistical packages such as SPSS, SAS, and R. Although applied less frequently, SEMs provide an additional viable alternative for conducting GT analyses.

GT SEM for a $p \times i$ Design with Four Items



GT SEM for a $p \times i \times o$ Design with Four Items and Two Occasions

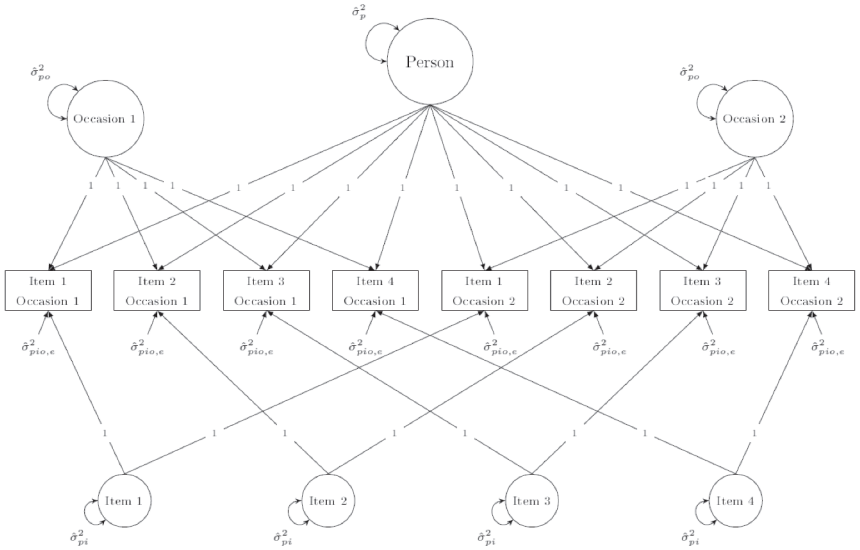


Figure 1. SEMs for one and two Facet GT Designs. GT SEM for a $p \times i$ Design with Four Items. GT SEM for a $p \times i \times o$ Design with Four Items and Two Occasions.

The top diagram in Figure 1 represents an SEM for a $p \times i$ design that can be used to estimate variance components relevant to norm-referencing purposes. The SEM has a person factor and separate orthogonal factors for each item. Factor loadings are constrained to equal one, and uniquenesses are set equal, thereby creating a model in which only two variance components are directly estimated: person variance ($\hat{\sigma}_p^2$) and overall measurement error variance ($\hat{\sigma}_{pi,e}^2$). These variance components can then be inserted into Equation 3 to compute a generalizability ($\hat{G}$ or $E\hat{\rho}^2$) coefficient.

$$\hat{G}\text{ coefficient for } p \times I \text{ design} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \left(\frac{\hat{\sigma}_{pi,e}^2}{n_i'}\right)} \tag{3}$$

G coefficients mirror conventional reliability coefficients in that they reflect relative differences in scores used for norm-referencing purposes (e.g., rank ordering). The variance component for items ($\hat{\sigma}_i^2$) is missing from Equation 3 because mean levels of item scores are constants that do not affect relative differences between persons. However, differences in item means would be relevant when making criterion-referenced decisions based on absolute levels of scores. Global and cut-score specific dependability (D or ϕ) coefficients shown in Equations 4 and 5 can be used for those purposes because

they take both relative and absolute differences in scores into account (Brennan, 2001; Brennan & Kane, 1977; Kane & Brennan, 1980).

$$\text{Global } \hat{\Phi} \text{ coefficient for } p \times I \text{ design} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \left(\frac{\hat{\sigma}_{pi,e}^2}{n_i'} + \frac{\hat{\sigma}_i^2}{n_i'}\right)} \tag{4}$$

$$\text{Cut-score specific } \hat{\Phi} \text{ coefficient} = \frac{\hat{\sigma}_p^2 + \left[(\bar{Y} - \text{Cut score})^2 - \hat{\sigma}_{\bar{Y}}^2\right]}{\hat{\sigma}_p^2 + \left[(\bar{Y} - \text{Cut score})^2 - \hat{\sigma}_{\bar{Y}}^2\right] + \left(\frac{\hat{\sigma}_{pi,e}^2}{n_i'} + \frac{\hat{\sigma}_i^2}{n_i'}\right)}, \tag{5}$$

where $\hat{\sigma}_{\bar{Y}}^2 = \frac{\hat{\sigma}_p^2}{n_p'} + \frac{\hat{\sigma}_{pi,e}^2}{n_p' n_i'} + \frac{\hat{\sigma}_i^2}{n_i'}$ and corrects for bias (see Brennan & Kane, 1977).

Terms within parentheses in the denominators of Equations 3 and 4 represent *relative error* and *absolute error*, respectively. When item means are equal (i.e., $\hat{\sigma}_i^2 = 0$), relative and absolute error coincide as do G and global D coefficients. As a result, differences between G and global D coefficients reveal the extent to which item mean differences lower the overall dependability of scores. When decisions are made based on targeted cut points along the observed score scale, cut-score specific D coefficients are particularly useful because they reflect levels of agreement in classifications over random repetitions of the assessment procedure (Kane & Brennan, 1980).

Jorgensen (2021) notes that, while the variance component for items is not directly obtainable from the factor model in Figure 1, it is a function of mean structure parameters when they are constrained to match GT definitions of mean deviation scores in which the Person factor's mean equals the grand mean across all measurement conditions, and intercepts represent mean deviation scores. Such constraints mirror those imposed when identifying SEMs using effect coding rather than the common marker method in which the intercept for one of the indicators for a given factor is set equal to zero and the loading for that indicator is set equal to one. With effect coding (see Little, Siegers, & Card, 2006), item intercepts are constrained to sum to zero, and factor loadings are constrained to average one (or equivalently sum to equal the number of items). Accordingly, mean deviation scores are distributed around zero (i.e., centered) with the same variance as the original item scores, and the variance component for items can be estimated using Equation 6.

$$\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{1}^{n_i} (Intercept_i)^2 \tag{6}$$

2.3. Two Facet Designs

The real advantages of GT over conventional measurement models emerge when we take multiple sources of measurement error into account. The *persons by items by occasions* ($p \times i \times o$) random-facets design is used with objectively scored measures to partition estimated error variance into categories reflecting inter-person differences in ordering of item scores (specific-factor error; $\hat{\sigma}_{pi}^2$), ordering of occasion scores (transient error, $\hat{\sigma}_{po}^2$), and within-occasion "noise" (random-response error, $\hat{\sigma}_{pio,e}^2$). "Specific-factor" is a term from the factor-analysis literature (e.g., Brown, 2006, *p.* 17) for a source of error that is "reliable" in the sense that it can be expected as a source of variance under any condition in which the item is administered. In the present context, such effects represent enduring idiosyncratic characteristics of items that are unrelated to the targeted construct(s) being measured and thus considered measurement error. Transient error represents person-specific effects on scores that are pervasive within the occasion of assessment but not across occasions. These effects might stem from test-taker/respondent dispositions, mind-sets, and physiological conditions; their reactions to administration and environmental factors; and other temporary entities within the assessment setting on whole that affect behavior. Random-response error corresponds to fleeting within-occasion effects on a person's scores that follow no systematic pattern

(e.g., momentary lapses in attention, fluctuations in moods, changes in motivation, distractions, etc.; see, e.g., Le, Schmidt, & Putka, 2009; Schmidt, Le, & Ilies, 2003; Thorndike, 1951). In frameworks such as latent state-trait theory, specific-factor and transient error would, respectively, be labeled as *method* and *state* effects (see, e.g., Vispoel, Xu, & Schneider, 2022a).

Equations 7 and 8 represent partitioning of variance at individual item and item-mean score levels for the $p \times i \times o$ design. Altogether, there are seven relevant variance components, including one for persons ($\hat{\sigma}_p^2$), three for the relative sources of measurement error related to persons just mentioned ($\hat{\sigma}_{pi}^2, \hat{\sigma}_{po}^2, \hat{\sigma}_{pio,e}^2$), and three for differences in absolute levels of mean scores ($\hat{\sigma}_i^2, \hat{\sigma}_o^2, \hat{\sigma}_{io}^2$). The letters I and O are capitalized in Equation 8 and elsewhere to signify averaging over item and occasion scores. Primes over *n*s allow for possible changes in numbers of items and/or occasions within decision studies.

$$p \times i \times o \text{ design individual item score level} : \hat{\sigma}_{Y_{pio}}^2 = \hat{\sigma}_p^2 + \hat{\sigma}_{pi}^2 + \hat{\sigma}_{po}^2 + \hat{\sigma}_{pio,e}^2 + \hat{\sigma}_i^2 + \hat{\sigma}_o^2 + \hat{\sigma}_{io}^2, \quad (7)$$

$$p \times I \times O \text{ design item} - \text{mean score level} : \hat{\sigma}_{Y_{pIO}}^2 = \hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi}^2}{n_i'} + \frac{\hat{\sigma}_{po}^2}{n_o'} + \frac{\hat{\sigma}_{pio,e}^2}{n_i' n_o'}, \quad (8)$$

where $\hat{\sigma}^2 =$ estimated variance component, $Y_{pio} =$ score for a particular person on a given combination of item and occasion, $Y_{pIO} =$ mean across all items and occasions for a particular person, $n_i' =$ number of items, and $n_o' =$ number of occasions.

The bottom diagram in Figure 1 represents a simple SEM for the $p \times i \times o$ design for deriving variance components relevant to norm-referencing purposes. The SEM has orthogonal factors for person, each item, and each occasion. Factor loadings are set equal to 1, and item variances, occasion variances, and item uniquenesses are, respectively, set equal. In all, four variance components are estimated that represent universe or person scores ($\hat{\sigma}_p^2$) and the three types of measurement error previously mentioned: specific-factor ($\hat{\sigma}_{pi}^2$), transient ($\hat{\sigma}_{po}^2$) and random-response ($\hat{\sigma}_{pio,e}^2$). We then can insert these estimates into Equation 9 to derive a G coefficient that takes all three sources of error into account.

$$\hat{\text{G}} \text{ coefficient for } p \times I \times O \text{ design} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \left(\frac{\hat{\sigma}_{pi}^2}{n_i'} + \frac{\hat{\sigma}_{po}^2}{n_o'} + \frac{\hat{\sigma}_{pio,e}^2}{n_i' n_o'}\right)} \quad (9)$$

However, as before, the variance components ($\hat{\sigma}_i^2, \hat{\sigma}_o^2, \hat{\sigma}_{io}^2$) needed for dependability coefficients are missing. To estimate the remaining variance components, we again place constraints on the SEM's mean structure (see Jorgensen, 2021). First, the sum of all item factor means is set equal to zero. Second, the sum of all occasion factor means is set to zero. Finally, the sum of intercepts for all indicators (combinations of items and occasions now) is set equal to zero. After imposing these constraints, we can estimate the remaining variance components for absolute error using Equations 10 to 12 and insert them into Equations 13 and 14 to compute global and cut-score specific D coefficients.

$$\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{1}^{n_i} (\textit{Item factor mean}_i)^2 \quad (10)$$

$$\hat{\sigma}_o^2 = \frac{1}{n_o - 1} \sum_{1}^{n_o} (\textit{Occasion factor mean}_o)^2 \quad (11)$$

$$\hat{\sigma}^2_{io} = \frac{1}{(n_i \times n_o) - 1} \sum_{1}^{n_i \times n_o} (Intercept_{io})^2 \tag{12}$$

$$\text{Global } \widehat{\text{D}} \text{ coefficient for } p \times I \times O \text{ design} = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \left(\frac{\hat{\sigma}^2_{pi}}{n'_i} + \frac{\hat{\sigma}^2_{po}}{n'_o} + \frac{\hat{\sigma}^2_{pio,e}}{n'_i n'_o} + \frac{\hat{\sigma}^2_i}{n'_i} + \frac{\hat{\sigma}^2_o}{n'_o} + \frac{\hat{\sigma}^2_{io}}{n'_i n'_o} \right)} \tag{13}$$

$$\text{Cut-score specific } \widehat{\text{D}} \text{ coefficient} = \frac{\hat{\sigma}^2_p + \left[\left(\overline{\text{Y}} - \text{Cut score} \right)^2 - \hat{\sigma}^2_{\overline{\text{Y}}} \right]}{\hat{\sigma}^2_p + \left[\left(\overline{\text{Y}} - \text{Cut score} \right)^2 - \hat{\sigma}^2_{\overline{\text{Y}}} \right] + \left(\frac{\hat{\sigma}^2_{pi}}{n'_i} + \frac{\hat{\sigma}^2_{po}}{n'_o} + \frac{\hat{\sigma}^2_{pio,e}}{n'_i n'_o} + \frac{\hat{\sigma}^2_i}{n'_i} + \frac{\hat{\sigma}^2_o}{n'_o} + \frac{\hat{\sigma}^2_{io}}{n'_i n'_o} \right)} \tag{14}$$

where $\hat{\sigma}^2_{\overline{\text{Y}}} = \frac{\hat{\sigma}^2_p}{n'_p} + \frac{\hat{\sigma}^2_{pi}}{n'_p n'_i} + \frac{\hat{\sigma}^2_{po}}{n'_p n'_o} + \frac{\hat{\sigma}^2_{pio,e}}{n'_p n'_i n'_o} + \frac{\hat{\sigma}^2_i}{n'_i} + \frac{\hat{\sigma}^2_o}{n'_o} + \frac{\hat{\sigma}^2_{io}}{n'_i n'_o}$ and corrects for bias (see Brennan & Kane, 1977).

2.4. Estimation Procedures

In traditional ANOVA applications of GT, variance components are estimated using expected mean squares. To reduce the occurrence of negative variance components, alternative estimation procedures such as maximum likelihood and restricted maximum likelihood are available in many packages. These estimation procedures typically provide highly similar results, but idiosyncrasies within algorithms used across packages and procedures may lead to differences, especially when negative variances are encountered (see, e.g., Marcoulides, 1990). To alert readers to such possible discrepancies, we used a variety of estimation procedures available within the GT variance component and SEM programs we compared.

3. This Investigation

Our initial goal for this investigation was accomplished in previous sections by showing how all variance components for one and two facet GT designs can be estimated using SEMs. Our further goals in sections to follow are: (1) to apply Jorgensen's procedures for estimating absolute error to data collected from a large sample of respondents in a live assessment setting, (2) to compare those results using ANOVA and SEM methods with a variety of packages and estimation procedures, and (3) to illustrate common applications of GT using either method.

4. Methods

4.1. Participants, Measures, and Procedure

We collected data from 821 college students from a large Midwestern university (76.84% female, 93.21% Caucasian, mean age = 21.54) who completed online versions of the Self-Perception Profile for College Students (SPPCS; Neemann & Harter, 2012) on two occasions, separated by a week. The governing institutional review board approved the study beforehand (ID# 200809738), and all students gave informed consent before participating. We chose the SPPCS for illustration here because it measures multiple dimensions of self-concept targeted specifically to college students.

4.1.1. SPPCS

Neemann and Harter (2012) created the SPPCS to assess traditional-aged full-time college students' self-perceived competence across multiple areas of functioning. The SPPCS has 54 items with one 6-item subscale to measure overall self-esteem (Global Self-Worth) and twelve additional 4-item

subscales to measure the following domain-specific aspects of self-concept: Scholastic Competence, Intellectual Ability, Job Competence, Creativity, Sense of Humor, Morality, Social Acceptance, Romantic Relationships, Close Friendships, Parent Relationships, Physical Appearance, and Athletic Competence. Respondents answer items using a 4-point structured alternative format intended to reduce socially desirable responding. Each item consists of a pair of conflicting descriptions in which respondents must first decide which statement better describes them and then decide whether that statement is *really true* or *sort of true*. *Really true* and *sort of true* responses to the statement more indicative of higher self-regard are, respectively, scored as 4 and 3, whereas *sort of true* and *really true* responses to the statement reflecting lower self-regard are scored, respectively, as 2 and 1. Evidence supporting the reliability and validity of SPPCS domain-specific subscale scores provided in its most recent manual (Neemann & Harter, 2012) includes alpha reliability estimates ranging from .76 to .96, exploratory factor analyses verifying that subscales measure distinguishable constructs, and correlation coefficients reflecting logically consistent relationships with each other, Self-Worth scores, and other external criteria. Although Neemann and Harter did not provide alpha reliability estimates for the Self-Worth subscale in their manual, Vispoel, Morris, and Sun (2019) reported that they ranged from .87 to .89 in a study of college students supporting interchangeability of SPPCS scores across four modes of paper and computer administration.

4.2. Analyses

We ran GT analyses for fully crossed $p \times i$ and $p \times i \times o$ random-facet designs for each subscale from the SPPCS. We then derived estimated variance components, G coefficients, and D coefficients using the following programs and estimation procedures: (a) the GT package GENOVA with ANOVA-based expected mean square (i.e., unweighted least squares) estimates (Crick & Brennan, 1983), (b) the *gtheory* package in R (Moore, 2016) with restricted maximum likelihood (REML) estimates obtained from *lme4* (Bates, Maechler, & Bolker, 2023), (c) the variance components package in SPSS with unweighted least squares (ULS), maximum likelihood (ML), and REML estimates, (d) the variance components package (PROC VARCOMP) in SAS with ULS, ML, and REML estimates, (e) the *car* package in R with ULS estimates (Fox, Weisberg, & Price, 2023), (f) the *lme4* package in R with ML and REML estimates (Bates, Maechler, & Bolker, 2023), and (g) the *lavaan* SEM package in R with ULS and ML estimates (Rosseel, 2012; Rosseel, Jorgensen, & Rockwood, 2023).

5. Results

5.1. Descriptive Statistics

Means, standard deviations, alpha coefficients, and test–retest coefficients appear in Table 1 for all scales. Overall, values for these indices are in line with those from studies previously cited.

5.2. Indices for Single Facet GT Designs

Given the large numbers of scales, estimation procedures, and computer packages represented in our analyses, we limit reported results here primarily to G and global D coefficients. Tables 2 and 3 respectively include those coefficients for the $p \times I$ design. Across all 13 scales, 3 estimation methods, and 13 procedures within GENOVA, R, SPSS, and SAS, the maximum difference between G coefficients equals 0.001 and the maximum difference between global D coefficients equals 0.007. For the same estimation method (ULS, ML, or REML) in the non-SEM analyses for SAS and R, G and global D coefficients are identical to the three decimal places shown in the tables. Coefficients for SPSS in the non-SEM analyses based on a common estimation method varied by no more than 0.001 from those for SAS and R. The largest differences between coefficients in the non-SEM analyses within Tables 2 and 3 relate more to estimation procedure than to the package used, with global D coefficients

Table 1. Descriptive statistics and conventional reliability estimates for SPPCS scores.

	Index for Time 1			Index for Time 2			
Subscale	Mean: Scale (Item)	SD: Scale (Item)	Alpha	Mean: Scale (Item)	SD: Scale (Item)	Alpha	Test-Retest
Scholastic Competence	11.26 (2.81)	2.50 (.62)	.819	11.54 (2.88)	2.41 (.60)	.828	.823
Intellectual Ability	12.08 (3.02)	2.67 (.67)	.872	12.12 (3.03)	2.69 (.67)	.875	.835
Job Competence	12.89 (3.22)	2.25 (.56)	.733	12.95 (3.24)	2.23 (.56)	.768	.698
Creativity	11.18 (2.80)	2.94 (.73)	.910	11.28 (2.82)	2.87 (.72)	.914	.838
Humor	13.75 (3.44)	2.32 (.58)	.855	13.68 (3.42)	2.40 (.60)	.874	.813
Morality	13.10 (3.28)	2.61 (.65)	.866	13.11 (3.28)	2.59 (.65)	.880	.798
Social Acceptance	12.56 (3.14)	2.88 (.72)	.853	12.61 (3.15)	2.81 (.70)	.858	.832
Romantic Relationships	10.96 (2.74)	3.39 (.85)	.895	11.12 (2.78)	3.38 (.84)	.912	.849
Close Friendships	13.34 (3.34)	2.92 (.73)	.868	13.45 (3.36)	2.90 (.72)	.889	.835
Parent Relationships	14.28 (3.57)	2.48 (.62)	.870	14.28 (3.57)	2.53 (.63)	.904	.859
Appearance	10.24 (2.56)	3.08 (.77)	.887	10.42 (2.60)	3.16 (.79)	.907	.864
Athletic Competence	10.87 (2.72)	3.55 (.89)	.930	10.99 (2.75)	3.49 (.87)	.935	.895
Global Self-Worth	19.06 (3.18)	3.66 (.61)	.896	19.08 (3.18)	3.64 (.61)	.900	.862
Mean	12.74 (3.06)	2.86 (.69)	.866	12.82 (3.08)	2.85(.69)	.880	.831

for ML results for the SPPCS Intellectual Ability subscale varying the most (0.007) with those for ULS and REML. This same pattern for largest difference (0.013) also holds for the ML variance component for items within that subscale (see Table 4). G and global D coefficients for ULS and ML from the R *lavaan* SEM analyses are more aligned with each other, varying at most by 0.001.

5.3. Indices for Two-Facet GT Designs

Similar patterns of relationships hold for G and global D coefficients within the $p \times I \times O$ designs shown in Tables 5 and 6. Across scales, estimation procedures, and packages, G coefficients vary by no more than 0.001, and global D coefficients by no more than 0.005. The largest differences again are between ML and other estimation procedures in the non-SEM analyses for global D coefficients and variance components for the SPPCS Intellectual Ability subscale (see Tables 4 & 6). As before, results for ULS and ML are closer to each other in the SEM than in the non-SEM analyses. In comparison to G and global D coefficients for the $p \times I$ designs, those for the $p \times I \times O$ designs are uniformly lower because they account for additional sources of measurement error.

5.4. Practical Applications of the GT Analyses

In this section, we illustrate selected applications of the present designs and direct readers to Brennan (2001), Vispoel, Hong, and Lee (2023); Vispoel, Morris, & Kilinc (2018a, 2018b, 2018c, 2018d, 2019), and Vispoel and Tao (2013) for additional ones. Although variance components from any GT analysis and subscale could be used, we limit examples to ULS estimates from the R *lavaan* analysis for SPPCS Scholastic Competence scores. In these examples, the variance components for $\hat{\sigma}^2_p$, $\hat{\sigma}^2_{pi,e}$, and $\hat{\sigma}^2_i$ within the $p \times i$ design, respectively, equal 0.319, 0.283, 0.015, and those for $\hat{\sigma}^2_p$,$\hat{\sigma}^2_{pi}$,$\hat{\sigma}^2_{po}$,$\hat{\sigma}^2_{pio,e}$,$\hat{\sigma}^2_i$,$\hat{\sigma}^2_o$, and $\hat{\sigma}^2_{io}$ within the $p \times i \times o$ design, respectively, equal 0.290, 0.082, 0.021, 0.186, 0.013, 0.002, 0.000 (see pages 17–26 of the online Supplemental Material).

5.4.1. Using Cut-Score Specific D Coefficients

Cut-score specific D coefficients (Brennan & Kane, 1977; Kane & Brennan, 1980) are unique to GT but less frequently reported than G or global D coefficients. They gauge levels of agreement in categorizing scores above or below targeted points along the assessment continuum when making criterion referenced decisions. Such decisions are common when using cut scores for screening, selection, classification, or domain referencing purposes. As with

Table 2. $p \times I$ design G coefficients for all SPPCS subscales, computer packages, and estimation procedures.

	Package/Estimation Procedure												
Subscale	GENOVA	R *gtheory*	SPSS VC ULS	SPSS VC ML	SPSS VC REML	SAS VC ULS	SAS VC ML	SAS VC REML	R VC ULS	R VC ML	R VC REML	*lavaan* ULS	*lavaan* ML
Scholastic Competence	.818	*.819*	.818	.818	.818	*.819*	*.819*	*.819*	*.819*	*.819*	*.819*	.818	.818
Intellectual Ability	.872	.872	.872	.872	.872	.872	.872	.872	.872	.872	.872	.872	.872
Job Competence	.733	.733	.733	.733	.733	.733	.733	.733	.733	.733	.733	.733	*.732*
Creativity	.910	.910	.910	.910	.910	.910	.910	.910	.910	.910	.910	.910	.910
Humor	.855	.855	.855	.855	.855	.855	.855	.855	.855	.855	.855	.855	.855
Morality	.866	.866	.866	.866	.866	.866	.866	.866	.866	.866	.866	.866	*.867*
Social Acceptance	.853	.853	.853	.853	.853	.853	.853	.853	.853	.853	.853	.853	.853
Romantic Relationships	.895	.895	.895	.895	.895	.895	.895	.895	.895	.895	.895	.895	.895
Close Friendships	.868	.868	.868	.868	.868	*.869*	.868	.868	.868	.868	.868	.868	.868
Parent Relationships	.871	*.870*	.871	.871	.871	*.870*	*.870*	*.870*	*.870*	*.870*	*.870*	.871	.871
Appearance	.887	.887	.887	.887	.887	.887	.887	.887	.887	.887	.887	.887	.887
Athletic Competence	.930	.930	.930	.930	.930	.930	.930	.930	.930	.930	.930	.930	.930
Global Self-Worth	.896	.896	.896	.896	.896	.896	.896	.896	.896	.896	.896	.896	.896
Mean	.866	.866	.866	.866	.866	.866	.866	.866	.866	.866	.866	.866	.866

VC= variance component program, ULS = unweighted least squares, ML = maximum likelihood, REML = restricted maximum likelihood. Italicized values in the body of the table represent ones that differ with corresponding values for *lavaan* ULS estimates.

G and global D coefficients, cut-score specific D coefficients can range from 0 to 1, with higher values reflecting greater levels of decision consistency. In the present examples, cut-score specific D coefficients are originally expressed on the item-mean metric (i.e., assuming a composite is calculated by averaging across item scores) but can be converted to a scale's total-score metric (i.e., assuming a composite is calculated by summing item scores) by multiplying item-mean scores by the number of items in the scale.

In Equations 15 and 16, we illustrate computation of cut-score specific D coefficients for the SPPCS Scholastic Competence scale for an arbitrarily chosen item-mean score of 3 for a person (i.e., total score of 12) within the $p \times I$ and $p \times I \times O$ designs based on $n'_i = 4$ and $n'_o = 1$. $\bar{Y}$s in the equations represent means for persons across all relevant facet conditions within the GT design of interest. In practice, decision makers using the SPPCS or another relevant measure would likely choose a less arbitrary cut score based on how that measure is being used. For example, to screen the top 5% of applicants within a degree-granting program, a cut score might be chosen associated with the 95th percentile among applicants or from a large-scale study used to establish population norms.

Note from Equations 15 and 16 that both cut-score specific coefficients for the item-mean score 3 exceed the values of their associated global counterparts reported earlier in Tables 3 and 6 (i.e., 0.825 vs 0.811 in the $p \times I$ design and 0.768 vs 0.757 in the $p \times I \times O$ design). This relationship will generally hold for all scores that deviate from the scale mean (see Figure 2).

$\widehat{\mathrm{D}}$ coefficient with in the $p \times \mathrm{I}$ design for a Scholastic Competence item-mean score of 3

$$
\begin{aligned}
&= \frac{\widehat{\sigma}^2_p + \left[\left(\bar{\mathrm{Y}} - \text{Cut score}\right)^2 - \widehat{\sigma}^2_{\bar{\mathrm{Y}}}\right]}{\widehat{\sigma}^2_p + \left[\left(\bar{\mathrm{Y}} - \text{Cut score}\right)^2 - \widehat{\sigma}^2_{\bar{\mathrm{Y}}}\right] + \left(\frac{\widehat{\sigma}^2_{pi,e}}{n'_i} + \frac{\widehat{\sigma}^2_i}{n'_i}\right)} \\
&= \frac{.319 + \left[(2.81 - 3)^2 - .004225\right]}{.319 + \left[(2.81 - 3)^2 - .004225\right] + \left(\frac{.283}{4} + \frac{.015}{4}\right)} = .825, \\
&\text{where } \hat{\sigma}^2_{\bar{\mathrm{Y}}} = \frac{\hat{\sigma}^2_p}{n'_p} + \frac{\hat{\sigma}^2_{pi,e}}{n'_p n'_i} + \frac{\hat{\sigma}^2_i}{n'_i} = \frac{.319}{821} + \frac{.283}{821(4)} + \frac{.015}{4} = .004225
\end{aligned}
\tag{15}
$$

$\hat{\mathrm{D}}$ coefficient within the $p \times I \times O$ design for a Scholastic Competence item-mean score of 3

$$
\begin{aligned}
&= \frac{\widehat{\sigma}^2_p + \left[\left(\bar{\mathrm{Y}} - \text{Cut score}\right)^2 - \widehat{\sigma}^2_{\bar{\mathrm{Y}}}\right]}{\widehat{\sigma}^2_p + \left[\left(\bar{\mathrm{Y}} - \text{Cut score}\right)^2 - \widehat{\sigma}^2_{\bar{\mathrm{Y}}}\right] + \left(\frac{\widehat{\sigma}^2_{pi}}{n'_i} + \frac{\widehat{\sigma}^2_{po}}{n'_o} + \frac{\widehat{\sigma}^2_{pio,e}}{n'_i n'_o} + \frac{\widehat{\sigma}^2_i}{n'_i} + \frac{\widehat{\sigma}^2_o}{n'_o} + \frac{\widehat{\sigma}^2_{io}}{n'_i n'_o}\right)} \\
&= \frac{.290 + \left[(2.845 - 3)^2 - .00571\right]}{.290 + \left[(2.845 - 3)^2 - .00571\right] + \left(\frac{.082}{4} + \frac{.021}{1} + \frac{.186}{4(1)} + \frac{.013}{4} + \frac{.002}{1} + \frac{.000}{4(1)}\right)} = .768, \text{ where } \hat{\sigma}^2_{\bar{\mathrm{Y}}} \\
&= \frac{\hat{\sigma}^2_p}{n'_p} + \frac{\hat{\sigma}^2_{pi}}{n'_p n'_i} + \frac{\hat{\sigma}^2_{po}}{n'_p n'_o} + \frac{\hat{\sigma}^2_{pio,e}}{n'_p n'_i n'_o} + \frac{\hat{\sigma}^2_i}{n'_i} + \frac{\hat{\sigma}^2_o}{n'_o} + \frac{\hat{\sigma}^2_{io}}{n'_i n'_o} = \frac{.290}{821} + \frac{.082}{821(4)} + \frac{.021}{821(1)} + \frac{.186}{821(4)(1)} \\
&+ \frac{.013}{4} + \frac{.002}{1} + \frac{.000}{4(1)} = .00571
\end{aligned}
\tag{16}
$$

In Figure 2 we provide cut-score specific D coefficients for all possible Scholastic Competence scores for the $p \times I$ and $p \times I \times O$ designs on both the item-mean and total-score metrics. Note that values for the coefficients are lowest around the scale means and increase as scores deviate away from those means, thereby demonstrating greater agreement as scores move toward the scale extremes. As is the case with G and global D coefficients, cut-score specific D coefficients are uniformly lower in the $p \times I \times O$ than in the $p \times I$ design

Table 3. $p \times I$ design global D coefficients for all SPPCS subscales, computer packages, and estimation procedures.

	Package/Estimation Procedure												
Subscale	GENOVA	R *gtheory*	SPSS VC ULS	SPSS VC ML	SPSS VC REML	SAS VC ULS	SAS VC ML	SAS VC REML	R VC ULS	R VC ML	R VC REML	*lavaan* ULS	*lavaan* ML
Scholastic Competence	.811	.811	.811	*.813*	.811	.811	*.813*	.811	.811	*.813*	.811	.811	.811
Intellectual Ability	.847	*.846*	.847	*.853*	.847	*.846*	*.853*	*.846*	*.846*	*.853*	*.846*	.847	.847
Job Competence	.728	.728	.728	*.730*	.728	.728	*.729*	.728	.728	*.729*	.728	.728	.728
Creativity	.907	.907	.907	.907	.907	.907	.907	.907	.907	.907	.907	.907	.907
Humor	.851	.851	.851	*.852*	.851	.851	*.852*	.851	.851	*.852*	.851	.851	.851
Morality	*.861*	*.861*	*.861*	*.862*	*.861*	*.861*	*.862*	*.861*	*.861*	*.862*	*.861*	.860	*.861*
Social Acceptance	.850	*.851*	.850	*.851*	.850	*.851*	*.851*	*.851*	*.851*	*.851*	*.851*	.850	.850
Romantic Relationships	.887	.887	.887	*.889*	.887	.887	*.888*	.887	.887	*.888*	.887	.887	*.886*
Close Friendships	.861	.861	.861	*.863*	.861	*.862*	*.863*	.861	.861	*.863*	.861	.861	.861
Parent Relationships	.869	*.870*	.869	*.870*	.869	*.870*	*.870*	*.870*	*.870*	*.870*	*.870*	.869	.869
Appearance	.871	.871	.871	*.875*	.871	.871	*.875*	.871	.871	*.875*	.871	.871	.871
Athletic Competence	.924	.924	.924	*.925*	.924	.924	*.925*	.924	.924	*.925*	.924	.924	.924
Global Self-Worth	.879	.879	.879	*.882*	.879	.879	*.882*	.879	.879	*.882*	.879	.879	.879
Mean	.857	.857	.857	*.859*	.857	.857	*.859*	.857	.857	*.859*	.857	.857	.857

VC= variance component program, ULS = unweighted least squares, ML = maximum likelihood, REML = restricted maximum likelihood. Italicized values in the body of the table represent ones that differ with corresponding values for *lavaan* ULS estimates.

because they take additional sources of measurement error into account. Although not reported here, similar patterns of relationships would hold for all subscales.

5.4.2. Using Proportions of Measurement Error

Proportions of observed score variance accounted for by individual sources of measurement error in the $p \times I \times O$ design can be estimated by replacing the numerator of the G coefficient formula with the index representing the relevant source of error from the denominator of that formula. We illustrate estimation of proportions of specific-factor, transient, and random-response measurement error for the Scholastic Competence scale for $n'_i = 4$ and $n'_o = 1$ in Equations 17–19.

$$\widehat{\text{Proportion}} \text{ of specific-factor error} = \frac{\frac{\hat{\sigma}^2_{pi}}{n'_i}}{\hat{\sigma}^2_p + \left(\frac{\hat{\sigma}^2_{pi}}{n'_i} + \frac{\hat{\sigma}^2_{po}}{n'_o} + \frac{\hat{\sigma}^2_{pio,e}}{n'_i n'_o}\right)} = \frac{\frac{.082}{4}}{.290 + \left(\frac{.082}{4} + \frac{.021}{1} + \frac{.186}{4(1)}\right)} = .054 \tag{17}$$

$$\widehat{\text{Proportion}} \text{ of transient error} = \frac{\frac{\hat{\sigma}^2_{po}}{n'_o}}{\hat{\sigma}^2_p + \left(\frac{\hat{\sigma}^2_{pi}}{n'_i} + \frac{\hat{\sigma}^2_{po}}{n'_o} + \frac{\hat{\sigma}^2_{pio,e}}{n'_i n'_o}\right)} = \frac{\frac{.021}{1}}{.290 + \left(\frac{.082}{4} + \frac{.021}{1} + \frac{.186}{4(1)}\right)} = .056 \tag{18}$$

$$\widehat{\text{Proportion}} \text{ random-response error} = \frac{\frac{\hat{\sigma}^2_{pio,e}}{n'_i n'_o}}{\hat{\sigma}^2_p + \left(\frac{\hat{\sigma}^2_{pi}}{n'_i} + \frac{\hat{\sigma}^2_{po}}{n'_o} + \frac{\hat{\sigma}^2_{pio,e}}{n'_i n'_o}\right)} = \frac{\frac{.186}{4(1)}}{.290 + \left(\frac{.082}{4} + \frac{.021}{1} + \frac{.186}{4(1)}\right)} = .123 \tag{19}$$

Note that each source of measurement error accounts for a non-trivial proportion of observed score variance ranging from 0.054 to 0.123. However, what would be considered a large or meaningful proportion of error would necessarily depend on the context in which scores are being used and the consequences of decisions made from those scores. In addition to yielding insights into their separate and combined effects, proportions of measurement error provide a useful basis for determining the best ways to alter measurement procedures to enhance score consistency. The most direct way to reduce a particular source of measurement error is to increase the number of conditions for its related facet. Because proportions of specific-factor and transient error for the Scholastic Competence scale do not vary much (0.054 vs 0.056), score consistency could be similarly enhanced by increasing items or occasions. However, such decisions can be made more precisely by directly estimating the magnitude of G coefficients, global D coefficients, and proportions of measurement error when making those changes.

5.4.3 Altering Numbers of Items and/or Occasions

Indices of score consistency and proportions of measurement error when changing numbers of conditions for facets can be easily estimated by inserting desired numbers of facet conditions into relevant formulas. In Equations 20–24, we illustrate how the G coefficient, global D coefficient, and proportions of measurement error change for the Scholastic Competence subscale when doubling the number of items and pooling results across two occasions.

$$\widehat{\text{G}} \text{ coefficient}\left(n'_i = 8, n'_o = 2\right) = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \left(\frac{\hat{\sigma}^2_{pi}}{n'_i} + \frac{\hat{\sigma}^2_{po}}{n'_o} + \frac{\hat{\sigma}^2_{pio,e}}{n'_i n'_o}\right)} = \frac{.290}{.290 + \left(\frac{.082}{8} + \frac{.021}{2} + \frac{.186}{8(2)}\right)} = .900 \tag{20}$$

Table 4. Variance components and indices of score consistency and measurement error for the SPPCS intellectual ability subscale.

		Index for $p \times i$ design					Index for $p \times i \times o$ design												
Package	Estimation Program	*p*	*pi,e*	*i*	G-coef (*i*)	D-coef (*i*)	*p*	*pi*	*po*	*pio,e*	*i*	*o*	*io*	G-coef (*i*)	G-coef (*o*)	G-coef (*io*)	D-coef (*io*)	SFE	TE	RRE
GENOVA		.389	.228	.054	.872	*.846*	.360	.059	.032	.168	.047	.000	*.001*	.874	.835	.802	.782	.033	*.072*	.094
R	*gtheory* package (REML)	.389	.228	.054	.872	*.846*	.360	.059	.032	.168	.047	.000	*.001*	*.873*	.835	.802	*.781*	.033	.071	.094
SPSS	Variance Components (ULS)	.389	.228	.054	.872	.847	.360	.059	.032	.168	.047	.000	*.001*	.874	.835	.802	*.781*	.033	.071	.094
	Variance Components (ML)	.389	.228	*.041*	.872	*.853*	.360	.059	.032	.168	*.035*	.000	*.001*	.874	.835	.802	*.786*	.033	.071	.094
	Variance Components (REML)	.389	.228	.054	.872	.847	.360	.059	.032	.168	.047	.000	*.001*	.874	.835	.802	*.781*	.033	.071	.094
SAS	Variance Components (ULS)	.389	.228	.054	.872	*.846*	.360	.059	.032	.168	.047	.000	*.001*	.874	.835	.802	.782	.033	*.072*	.094
	Variance Components (ML)	.389	.228	*.041*	.872	*.853*	.360	.059	.032	.168	*.035*	.000	*.001*	*.873*	.835	.802	*.786*	.033	.071	.094
	Variance Components (REML)	.389	.228	.054	.872	*.846*	.360	.059	.032	.168	.047	.000	*.001*	*.873*	.835	.802	*.781*	.033	.071	.094
R	Variance Components (ULS)	.389	.228	.054	.872	*.846*	.360	.059	.032	.168	.047	.000	*.001*	.874	.835	.802	.782	.033	*.072*	.094
	Variance Components (ML)	.389	.228	*.041*	.872	*.853*	.360	.059	.032	.168	*.035*	.000	*.001*	*.873*	.835	.802	*.786*	.033	.071	.094
	Variance Components (REML)	.389	.228	.054	.872	*.846*	.360	.059	.032	.168	.047	.000	*.001*	*.873*	.835	.802	*.781*	.033	.071	.094
R	*lavaan* (ULS)	.389	.228	.054	.872	.847	.360	.059	.032	.168	.047	.000	.000	.874	.835	.802	.782	.033	.071	.094
	lavaan (ML)	.389	.228	.054	.872	.847	.360	.059	.032	.168	.047	.000	.000	.874	.835	.802	.782	.033	.071	.094

ULS = unweighted least squares, ML = maximum likelihood, REML = restricted maximum likelihood, SFE = specific-factor error, TE = transient error, RRE = random-response error. G-coef(*i*) = G coefficient for just items, D-coef(*i*) = global D coefficient for just items, G-coef(*o*) = G coefficient for just occasions, G-coef(*io*) = G coefficient for both items and occasions, D-coef(*io*) = global D coefficient for both items and occasions. Italicized values in the body of the table represent ones that differ with corresponding values for *lavaan* ULS estimates.

Table 5. $p \times I \times O$ design G coefficients for all SPPCS subscales, computer packages, and estimation procedures.

Subscale	Package/Estimation Procedure												
	GENOVA	R *gtheory*	SPSS VC ULS	SPSS VC ML	SPSS VC REML	SAS VC ULS	SAS VC ML	SAS VC REML	R VC ULS	R VC ML	R VC REML	*lavaan* ULS	*lavaan* ML
Scholastic Competence	.767	*.768*	.767	*.768*	.767	*.768*	*.768*	*.768*	*.768*	*.768*	*.768*	.767	*.768*
Intellectual Ability	.802	.802	.802	.802	.802	.802	.802	.802	.802	.802	.802	.802	.802
Job Competence	.654	*.655*	.654	.654	.654	*.655*	*.655*	*.655*	*.655*	*.655*	*.655*	.654	.654
Creativity	*.820*	*.820*	.821	.821	.821	*.820*	*.820*	*.820*	*.820*	*.820*	*.820*	.821	.821
Humor	*.782*	*.782*	.781	.781	.781	*.782*	*.782*	*.782*	*.782*	*.782*	*.782*	.781	*.782*
Morality	.772	.772	.772	.772	.772	.772	.772	.772	.772	.772	.772	.772	*.771*
Social Acceptance	.790	*.791*	.790	.790	.790	*.791*	*.791*	*.791*	*.791*	*.791*	*.791*	.790	.790
Romantic Relationships	.815	.815	.815	.815	.815	.815	.815	.815	.815	.815	.815	.815	.815
Close Friendships	.799	.799	.799	.799	.799	.799	.799	.799	.799	.799	.799	.799	*.798*
Parent Relationships	.827	*.826*	.827	.827	.827	*.826*	*.826*	*.826*	*.826*	*.826*	*.826*	.827	.827
Appearance	.835	*.834*	.835	.835	.835	*.834*	*.834*	*.834*	*.834*	*.834*	*.834*	.835	.835
Athletic Competence	*.883*	*.883*	.882	.882	.882	*.883*	*.883*	*.883*	*.883*	*.883*	*.883*	.882	.882
Global Self-Worth	*.837*	.838	.838	.838	.838	*.837*	.838	.838	*.837*	.838	.838	.838	.838
Mean	.799	.799	.799	.799	.799	.799	.799	.799	.799	.799	.799	.799	.799

VC= variance component program, ULS = unweighted least squares, ML = maximum likelihood, REML = restricted maximum likelihood. Italicized values in the body of the table represent ones that differ with corresponding values for *lavaan* ULS estimates.

$$\text{Global } \widehat{D} \text{ coefficient}\left(n_i^{'}=8, n_o^{'}=2\right) = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \left(\frac{\hat{\sigma}_{pi}^2}{n_i^{'}} + \frac{\hat{\sigma}_{po}^2}{n_o^{'}} + \frac{\hat{\sigma}_{pio,e}^2}{n_i^{'}n_0^{'}} + \frac{\hat{\sigma}_i^2}{n_i^{'}} + \frac{\hat{\sigma}_o^2}{n_o^{'}} + \frac{\hat{\sigma}_{io}^2}{n_i^{'}n_o^{'}}\right)}$$
$$= \frac{.290}{.290 + \left(\frac{.082}{8} + \frac{.021}{2} + \frac{.186}{8(2)} + \frac{.013}{8} + \frac{.002}{2} + \frac{.000}{8(2)}\right)} = .892 \quad (21)$$

$$\widehat{\text{Proportion of specific-factor error}}\left(n_i^{'}=8, n_o^{'}=2\right) = \frac{\frac{\hat{\sigma}_{pi}^2}{n_i^{'}}}{\hat{\sigma}_p^2 + \left(\frac{\hat{\sigma}_{pi}^2}{n_i^{'}} + \frac{\hat{\sigma}_{po}^2}{n_o^{'}} + \frac{\hat{\sigma}_{pio,e}^2}{n_i^{'}n_o^{'}}\right)}$$
$$= \frac{\frac{.082}{8}}{.290 + \left(\frac{.082}{8} + \frac{.021}{2} + \frac{.186}{8(2)}\right)} = .032 \quad (22)$$

$$\widehat{\text{Proportion of transient error}}\left(n_i^{'}=8, n_o^{'}=2\right) = \frac{\frac{\hat{\sigma}_{po}^2}{n_o^{'}}}{\hat{\sigma}_p^2 + \left(\frac{\hat{\sigma}_{pi}^2}{n_i^{'}} + \frac{\hat{\sigma}_{po}^2}{n_o^{'}} + \frac{\hat{\sigma}_{pio,e}^2}{n_i^{'}n_o^{'}}\right)} = \frac{\frac{.021}{2}}{.290 + \left(\frac{.082}{8} + \frac{.021}{2} + \frac{.186}{8(2)}\right)}$$
$$= .033 \quad (23)$$

$$\widehat{\text{Proportion of random-response error}}\left(n_i^{'}=8, n_o^{'}=2\right) = \frac{\frac{\hat{\sigma}_{pio,e}^2}{n_i^{'}n_o^{'}}}{\hat{\sigma}_p^2 + \left(\frac{\hat{\sigma}_{pi}^2}{n_i^{'}} + \frac{\hat{\sigma}_{po}^2}{n_o^{'}} + \frac{\hat{\sigma}_{pio,e}^2}{n_i^{'}n_o^{'}}\right)}$$
$$= \frac{\frac{.186}{8(2)}}{.290 + \left(\frac{.082}{8} + \frac{.021}{2} + \frac{.186}{8(2)}\right)} = .036 \quad (24)$$

Note that values for G and global D coefficients increase, and proportions of measurement error decrease in relation to those originally reported in Tables 5 and 6 and Equations 17–19 in which numbers of items and occasions equal 4 and 1, respectively. More specifically, after doubling items and occasions, the G coefficient increases from 0.767 to 0.900 and global D coefficient from 0.757 to 0.892, whereas the proportion of specific-factor error drops from 0.054 to 0.032, transient error from 0.056 to 0.033, and random-response error from 0.123 to 0.036.

When deciding the best ways to change a measurement procedure, plots like those shown in Figure 3 are often created to estimate how score consistency is affected by a wide variety of changes to facet conditions. In the figure, we show G and global D coefficients for the Scholastic Competence subscale with number of items varying from 4 to 12 and number of occasions varying from 1 to 3. G and global D coefficients both improve with increases in numbers of items or occasions but to a diminishing extent with the same progressive incremental changes in facet conditions. Because proportions of specific-factor and transient errors are similar, doubling occasions or items would lead to comparable improvements in G and D coefficients. However, given a choice between the two, doubling items would likely be more practical, efficient, and cost-effective in the long run because only one administration of the measure would be required.

Table 6. $p \times I \times O$ design global D coefficients for all SPPCS subscales, computer packages, and estimation procedures.

	Package/Estimation Procedure												
Subscale	GENOVA	R *gtheory*	SPSS VC ULS	SPSS VC ML	SPSS VC REML	SAS VC ULS	SAS VC ML	SAS VC REML	R VC ULS	R VC ML	R VC REML	*lavaan* ULS	*lavaan* ML
Scholastic Competence	.757	.757	.757	*.759*	.757	.757	*.759*	.757	.757	*.759*	.757	.757	*.758*
Intellectual Ability	*.781*	*.781*	*.781*	*.786*	*.781*	.782	*.786*	*.781*	.782	*.786*	*.781*	.782	.782
Job Competence	.650	*.651*	.650	*.651*	.650	*.651*	*.652*	*.651*	*.651*	*.652*	*.651*	.650	.650
Creativity	.819	*.818*	.819	.819	.819	*.818*	*.818*	*.818*	*.818*	*.818*	*.818*	.819	.819
Humor	.778	*.779*	.778	*.779*	.778	*.779*	*.780*	*.779*	*.779*	*.780*	*.779*	.778	*.779*
Morality	.767	.767	.767	*.768*	.767	.767	*.768*	.767	.767	*.769*	.767	.767	.767
Social Acceptance	*.788*	*.789*	*.788*	*.788*	*.788*	*.789*	*.789*	*.789*	*.789*	*.789*	*.789*	.787	*.788*
Romantic Relationships	*.808*	*.808*	*.808*	*.809*	*.808*	*.808*	*.809*	*.808*	*.808*	*.809*	*.808*	.807	.807
Close Friendships	.793	.793	.793	*.794*	.793	.793	*.794*	.793	.793	*.794*	.793	.793	.793
Parent Relationships	.825	*.826*	.825	*.826*	.825	*.826*	*.826*	*.826*	*.826*	*.826*	*.826*	.825	.825
Appearance	.820	.820	.820	*.823*	.820	.820	*.823*	.820	.820	*.823*	.820	.820	.820
Athletic Competence	*.877*	.876	*.877*	*.878*	*.877*	.876	*.878*	.876	.876	*.878*	.876	.876	.876
Global Self-Worth	.825	*.824*	.825	*.827*	.825	.825	*.826*	*.824*	.825	*.826*	*.824*	.825	.825
Mean	.791	.791	.791	*.793*	.791	.791	*.793*	.791	.791	*.793*	.791	.791	.791

VC = variance component program, ULS = unweighted least squares, ML = maximum likelihood, REML = restricted maximum likelihood. Italicized values in the body of the table represent ones that differ with corresponding values for *lavaan* ULS estimates.

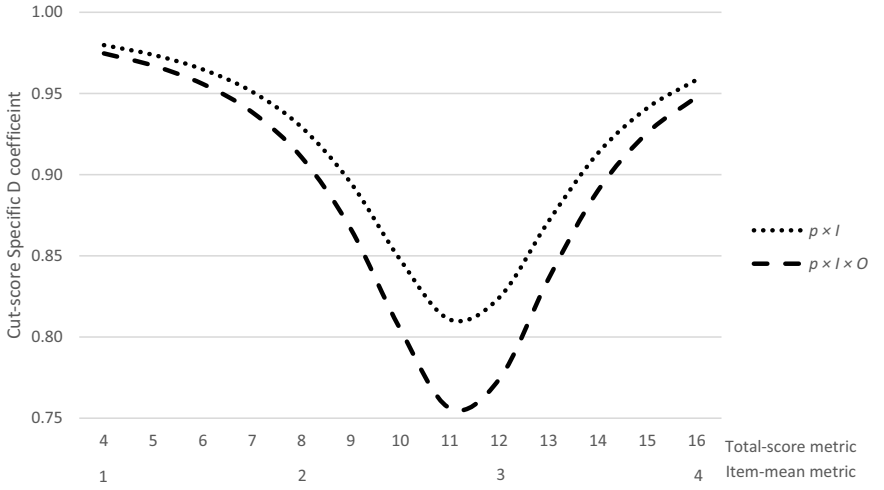


Figure 2. Cut-score specific D coefficients for the SPPCS scholastic competence subscale.

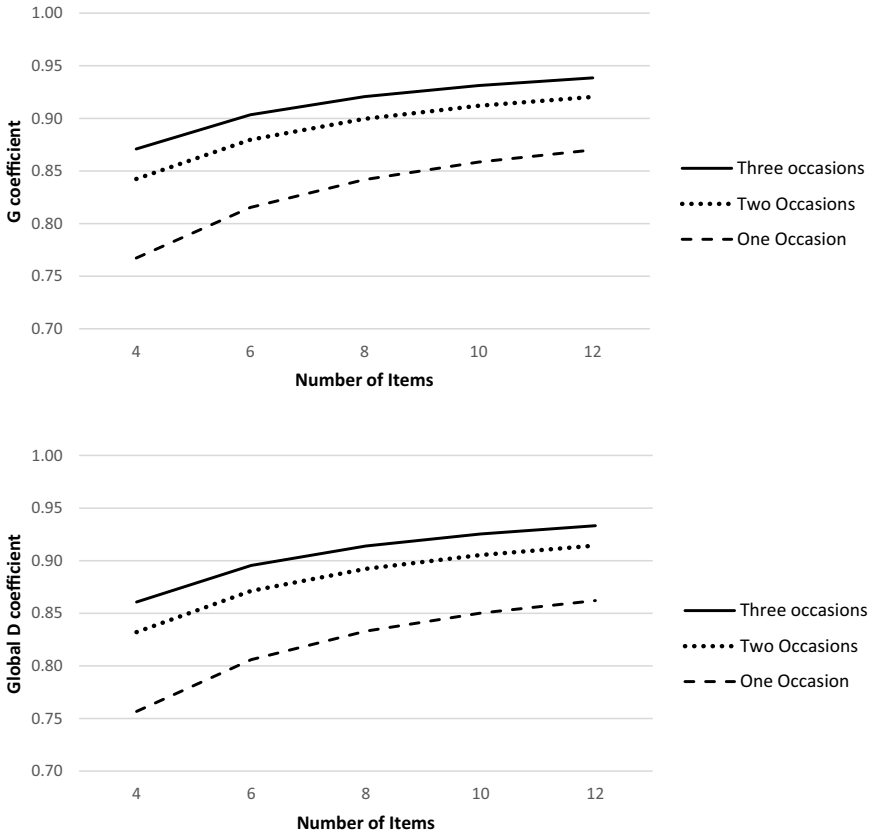


Figure 3. Estimated G and global D coefficients for the SPPCS scholastic competence subscale for differing numbers of items and occasions.

5.4.4. Restricting Universes of Generalization

Proportions of measurement error are also useful in gauging changes in reliability when restricting universes of generalization to a lesser number of facets. Such changes are analogous to treating the missing facet(s) as fixed. Percentages of upward change in reliability for such modifications would equal one hundred times the proportion of measurement error corresponding to the omitted facet(s) divided by the G coefficient that includes all sources of measurement error. When limiting the universe of generalization to just items and excluding transient effects as measurement error, the generalizability of Scholastic Competence scores increases by 7.30% ($100 \times 0.056/0.767$). Restricting universes in this manner would be appropriate when measuring states expected to change noticeably from one occasion to the next. When limiting the universe of generalization to just occasions and excluding specific-factor effects as measurement error, the generalizability of Scholastic Competence scores increases by 7.04% ($100 \times 0.054/0.767$). Such restrictions might be appropriate when universe score and enduring specific-factor or item method effects together increase the concurrent or predictive validity of observed scores.

G coefficients for more restricted universes of generalization can be derived by adding the proportion of measurement error variance for the excluded facet(s) to the G coefficient for the full design. In the present example, G coefficients would equal 0.823 (0.767 +0 .056) for generalizing across just items and 0.821 (0.767 + 0.054) for generalizing just across occasions. Both coefficients illustrate noteworthy increases in generalizability when a key source of measurement error is omitted.

6. Discussion

6.1. Overview

Marcoulides (1996) and Raykov and Marcoulides (2006) were among the first to illustrate how GT designs could be analyzed within SEM frameworks, and other researchers have recently revisited and extended those techniques (see, e.g., Ark, 2015; Jorgensen, 2021; Morris, 2020 Vispoel, Lee, Chen, & Hong, 2023a, 2023b, 2023c; Vispoel, Morris, & Kilinc, 2018a, 2018b; Vispoel, Lee, & Hong, 2023; Vispoel, Lee, Xu, & Hong, 2022, 2023; Vispoel, Xu, & Kilinc, 2021; Vispoel, Xu, & Schneider, 2022a). However, these applications overall emphasized indices reflecting relative differences among scores either because the researchers were only interested in those indices or considered derivation of those for absolute differences from SEMs unwieldy and/or subject to severe technical limitations. For example, Ark (2015) described a Q method for each facet of interest in which facet conditions represent rows and objects of measurement represent columns in the analyzed data matrix. Under these conditions, indices of absolute error could only be derived when the number of conditions for facets exceeds the number of objects of measurement, which would rarely be the case in practice.

To address this problem, Jorgensen (2021) used a small set of generated non-empirical data to demonstrate how variance components for absolute differences could be derived from the same SEMs used in previous studies by setting appropriate constraints on intercepts, factor loadings, and factor means. Our goals in the present study were (1) to illustrate these SEM procedures for analyzing complete GT designs, (2) to compare variance components and related indices from appropriately constrained SEMs within the *lavaan* program in R to those obtained from popular ANOVA-based packages using data from a large sample of respondents who completed measures in a live assessment setting, and (3) to describe practical applications of GT using either approach.

6.2. Congruence of Results Across Packages

As anticipated, results for G coefficients, D coefficients, and variance components across models, packages, and estimation procedures were very consistent, with G coefficients and global

D coefficients varying by no more than 0.001 and 0.007, respectively. Although ML, ULS, and REML estimates typically produced similar results, ML estimates for absolute error within the variance component programs in SPSS, SAS, and R sometimes differed with ULS and REML estimates within those programs and with ULS and ML estimates in the *lavaan* SEM analyses in R. Differences in ML estimates between the SEM and non-SEM programs are likely due to variations in discrepancy functions and divisors used when estimating variances. The sum-of-squares or negative log-likelihood function is minimized with respect to each row of data (i.e., observed vs predicted scores case by case) in the variance component programs, whereas differences between observed and predicted values within the covariance matrix are minimized in the SEM program. Also, in the variance component programs, n is used in the denominator for ML variance estimates for absolute error, whereas $n - 1$ is used for ULS and REML (Raudenbush & Bryk, 2002, *pp*. 52–53). For example, if we multiply the SPPSC Intellectual Ability subscale's variance component for items of 0.054 using ULS or REML in Table 4 by ¾, you get the value of 0.401 within rounding for the corresponding variance component for ML. As a result, variance component estimates for ML within the non-SEM programs have a slight downward bias, and D coefficient estimates have a slight upward bias that was not evident in the SEM analyses. Indices derived using ULS and REML varied negligibly across programs, but REML would have the advantage of eliminating potential negative variance components (Marcoulides, 1990).

Overall, these results from extended empirical data serve to verify that Jorgensen's methods applied to one and two facet GT-SEMs can yield all variance components and related indices for complete GT designs in line with those obtained from numerous other packages. As a result, researchers more familiar with structural modeling software need not rely on specialized GT or variance component programs to perform traditional GT analyses and feel more confident in the credibility of their findings. Although we focused on objectively scored measures of self-concept here, the techniques we describe are equally applicable to subjectively and objectively scored measures of achievement, aptitude, behavior, and psychomotor skills. As an example, a *persons by items by raters* ($p \times i \times r$) design could be analyzed using the same methods illustrated here by substituting holistically scored essays for items and raters for occasions. In that design, the *pi*, *pr*, and *pir,e* variance components would represent inter-task, inter-rater, and random scoring error, respectively.

6.3. Applying GT in Practice

Examples illustrated here were intended to show practical applications of GT when using either SEM or ANOVA-based procedures. Such applications included derivation of score consistency indices for both norm- and criterion-referencing purposes, partitioning observed score variance into components representing universe score and multiple sources of measurement error, and estimation of effects of possible changes made to measurement procedures. A consistent theme across these applications was the importance of taking multiple sources of measurement error into account and potential over-estimation of score consistency when relevant sources of error are excluded.

However, applications of GT-SEMs can go far beyond replicating variance components provided in more traditional ways. Examples of such extensions include:

(a) Deriving Monte Carlo-based confidence intervals to evaluate the trustworthiness of estimated variance components, G coefficients, D coefficients, and proportions of measurement error (Jorgensen, 2021; Jorgensen, Pornprasertmanit, Schoemann, & Rosseel, 2022; Vispoel, Hong, & Lee, 2023; Vispoel, Lee, Chen, & Hong, 2023a, 2023b, 2023c; Vispoel, Lee, & Hong, 2023; Vispoel, Lee, Hong, & Chen, 2023);
(b) Using alternative estimation procedures (e.g., diagonally weighted least squares) to correct for scale coarseness due to limited numbers of response options and/or unequal underlying intervals between those options (Ark, 2015; Jorgensen, 2021; Vispoel, Hong, & Lee, 2023; Vispoel, Lee, Chen, & Hong, 2023a; Vispoel, Lee, Xu, & Hong, 2023; Vispoel, Morris, & Kilinc, 2018a, 2019);

(c) Allowing for congeneric relationships between indicators and underlying factors (Vispoel, Hong, & Lee, 2023; Vispoel, Lee, Xu, & Hong, 2022; Vispoel, Xu, & Kilinc, 2021; Vispoel, Xu, & Schneider, 2022a);
(d) Extending partitioning of variance to the individual item level (Vispoel, Hong, & Lee, 2023; Vispoel, Lee, Xu, & Hong, 2022; Vispoel, Xu, & Kilinc, 2021; Vispoel, Xu, & Schneider, 2022a);
(e) Formally testing overall model fit when warranted (Vispoel, Hong, & Lee, 2023; Vispoel, Lee, Xu, & Hong, 2022; Vispoel, Morris, & Kilinc, 2018b, 2019; Vispoel, Xu, & Kilinc, 2021; Vispoel, Xu, & Schneider, 2022a, 2022b)
(f) Accounting for method factor effects due to item wording (Vispoel, Hong, & Lee, 2023; Vispoel, Xu, & Schneider, 2022a)
(g) Integrating analyses into latent state-trait theoretical frameworks (Vispoel, 2023b; Vispoel, Xu, & Kilinc, 2021; Vispoel, Xu, & Schneider, 2022a); and
(h) Expanding analyses to multivariate (Vispoel, Hong, & Lee, 2023; Vispoel, Lee, & Hong, 2023; Vispoel, Lee, Chen, & Hong, 2023c; Vispoel, Lee, Hong, & Chen, 2023) and bifactor model designs (Vispoel, Hong, & Lee, 2023; Vispoel & Lee, 2023; Vispoel, Lee, Chen, & Hong, 2023b, 2023c; Vispoel, Lee, Xu, & Hong 2022, 2023).

We refer readers to the studies cited above for further details and to our online Supplemental Material for examples and relevant code for performing the analyses demonstrated here using GENOVA, SPSS, SAS, and R.

Acknowledgments

We thank the Iowa Measurement Research Foundation for providing research assistant support for this project and Tingting Chen for her assistance in creating factor model diagrams and proof reading drafts of the article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the Iowa Measurement Research Foundation.

ORCID

Walter P. Vispoel http://orcid.org/0000-0002-9415-251X
Hyeri Hong http://orcid.org/0000-0002-7576-2574
Hyeryung Lee http://orcid.org/0000-0001-7642-6161
Terrence D. Jorgensen http://orcid.org/0000-0001-5111-6773

References

Ark, T. K. (2015). *Ordinal generalizability theory using an underlying latent variable framework* (Doctoral dissertation). University of British Columbia, Vancouver, British Columbia, Canada.
Bates, D., Maechler, M., & Bolker, B. (2023). Package 'lme4'. R package version (1.1-32). https://cran.r-project.org/web/packages/lme4/lme4.pdf
Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.
Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, *14*(3), 277–289. doi:10.1111/j.1745-3984.1977.tb00045.x
Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.
Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system (American College Testing Technical Bulletin 43)*. Iowa City: ACT, Inc.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.
Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163. doi:10.1111/j.2044-8317.1963.tb00206.x
Fox, J., Weisberg, S., & Price, B. (2023). Package 'car'. R package version (3.1-2). https://cran.r-project.org/web/packages/car/car.pdf
Jorgensen, T. D. (2021). How to estimate absolute-error components in structural equation models of generalizability theory. *Psych*, *3*(2), 113–133. doi:10.3390/psych3020011
Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful Tools for Structural Equation modeling*. R Package Version 0.5-6. https://CRAN.R-project.org/package=semTools
Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement*, *4*(1), 105–126. doi:10.1177/014662168000400111
Le, H., Schmidt, F. L., & Putka, D. J. (2009). The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organizational Research Methods*, *12*(1), 165–200. doi:10.1177/1094428107302900
Little, T. D., Siegers, D. W., & Card, A. (2006). A non-arbitrary method or identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*(1), 59–72. doi:10.1207/s15328007sem1301_3
Marcoulides, G. A. (1990). An alternative method for estimating variance components in generalizability theory. *Psychological Reports*, *66*(2), 379–386. doi:10.2466/pr0.1990.66.2.379
Marcoulides, G. A. (1996). Estimating variance components in generalizability theory: The covariance structure analysis approach [Teacher's corner]. *Structural Equation Modeling*, *3*(3), 290–299. doi:10.1080/10705519609540045
Moore, C. T. (2016). *gtheory: Apply Generalizability Theory with R. R Package Version .1.2*. https://CRAN.R-project.org/package=gtheory.
Morris, C. A. (2020). *Optimal methods for disattenuating correlation coefficients under realistic measurement conditions with single-form, self-report instruments* (Doctoral dissertation). University of Iowa. ProQuest Dissertation and Theses database.
Neemann, J., & Harter, S. (2012). *Self-Perception Profile for College Students: Manual and questionnaires (revised)*. Denver, CO: University of Denver.
Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
Raykov, T., & Marcoulides, G. A. (2006). Estimation of generalizability coefficients via a structural equation modeling approach to scale reliability evaluation. *International Journal of Testing*, *6*(1), 81–95. doi:10.1207/s15327574ijt0601_5
Rosseel, Y. (2012). *lavaan*: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi:10.18637/jss.v048.i02
Rosseel, Y., Jorgensen, T. D., & Rockwood, N. (2023). Package 'lavaan'. R package version (0.6-15). https://cran.r-project.org/web/packages/lavaan/lavaan.pdf
Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods*, *8*(2), 206–224. doi:10.1037/1082-989X.8.2.206
Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.
Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560–620). Washington, DC: American Council on Education.
Vispoel, W. P. (2023a). Why reliability for trait-based score is typically overestimated and what to do about it. *Biomedical Journal of Scientific & Technical Research*, *52*(4), 43883–42886. doi:10.26717/BJSTR.2023.52.008280
Vispoel, W. P. (2023b). Self-concept inventories measure more than just psychological traits. *Psychology Journal: Research Open*, *5*(4), 1–3. https://researchopenworld.com/self-concept-inventories-measure-more-than-just-psychological-traits/
Vispoel, W. P., Hong, H., & Lee, H. (2023). Benefits of doing generalizability theory analyses within structural equation modeling frameworks: Illustrations using the Rosenberg Self-Esteem Scale [Teacher's corner]. *Structural Equation Modeling: An Interdisciplinary Journal*, 1–17. Advance online publication. doi:10.1080/10705511.2023.2187734
Vispoel, W. P., & Lee, H. (2023). Merging generalizability theory and bifactor modeling to improve psychological assessments. *Psychology and Psychotherapy: Review Study*, *7*, 1–4. https://crimsonpublishers.com/pprs/pdf/PPRS.000652.pdf
Vispoel, W. P., Lee, H., Chen, T., & Hong, H. (2023a). Using structural equation modeling techniques to reproduce and extend ANOVA-based generalizability theory analyses for psychological assessments. *Psych*, *5*(2), 249–273. doi:10.3390/psych5020019
Vispoel, W. P., Lee, H., Chen, T., & Hong, H. (2023b). Extending applications of generalizability theory-based bifactor model designs. *Psych*, *5*(2), 545–575. doi:10.3390/psych5020036
Vispoel, W. P., Lee, H., Chen, T., & Hong, H. (2023c). Analyzing and comparing univariate, multivariate, and bifactor generalizability designs for hierarchically structured personality traits. *Journal of Personality Assessment*, 1–16. Advance online publication. doi:10.1080/00223891.2023.2268193

Vispoel, W. P., Lee, H., & Hong, H. (2023). Analyzing multivariate generalizability theory designs for psychological assessments within structural equation modeling frameworks [Teacher's corner]. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–19. Advance online publication. doi:10.1080/10705511.2023.2222913

Vispoel, W. P., Lee, H., Hong, H., & Chen, T. (2023). Applying multivariate generalizability theory to psychological assessments. *Psychological Methods*, 1–23. doi:10.1037/met0000606

Vispoel, W. P., Lee, H., Xu, G., & Hong, H. (2022). Expanding bifactor models of psychological traits to account for multiple sources of measurement error. *Psychological Assessment*, *32*(12), 1093–1111. doi:10.1037/pas0001170

Vispoel, W. P., Lee, H., Xu, G., & Hong, H. (2023). Integrating bifactor models into a generalizability theory based structural equation modeling framework. *The Journal of Experimental Education*, *91*(4), 718–738. doi:10.1080/00220973.2022.2092833

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018a). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, *23*(1), 1–26. doi:10.1037/met0000107

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018b). Practical applications of generalizability theory for designing, evaluating, and improving psychological assessments. *Journal of Personality Assessment*, *100*(1), 53–67. doi:10.1080/00223891.2017.1296455

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018c). Using generalizability theory to disattenuate correlation coefficients for multiple sources of measurement error. *Multivariate Behavioral Research*, *53*(4), 481–501. doi:10.1080/00273171.2018.1457938

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018d). Using G-theory to enhance evidence of reliability and validity for common uses of the Paulhus Deception Scales. *Assessment*, *25*(1), 69–83. doi:10.1177/1073191116641182

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2019). Using generalizability theory with continuous latent response variables. *Psychological Methods*, *24*(2), 153–178. doi:10.1037/met0000177

Vispoel, W. P., Morris, C. A., & Sun, L. (2019). Computerized and traditional administration of questionnaires: Psychometric quality and completion time for measures of self-concept. *The Journal of Experimental Education*, *87*(3), 384–399. doi:10.1080/00220973.2018.1448748

Vispoel, W. P., & Tao, S. (2013). A generalizability analysis of score consistency for the Balanced Inventory of Desirable Responding. *Psychological Assessment*, *25*(1), 94–104. doi:10.1037/a0029061

Vispoel, W. P., Xu, G., & Kilinc, M. (2021). Expanding G-theory models to incorporate congeneric relationships: Illustrations using the Big Five Inventory. *Journal of Personality Assessment*, *104*(4), 429–442. doi:10.1080/00223891.2020.1808474

Vispoel, W. P., Xu, G., & Schneider, W. S. (2022a). Interrelationships between latent state-trait theory and generalizability theory in a structural equation modeling framework. *Psychological Methods*, *27*(5), 773–803. doi:10.1037/met0000290

Vispoel, W. P., Xu, G., & Schneider, W. S. (2022b). Using parallel splits with self-report and other measures to enhance precision in generalizability theory analyses. *Journal of Personality Assessment*, *104*(3), 303–319. doi:10.1080/00223891.2021.1938589