AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

ISSN: 1082-989X

https://doi.org/10.1037/met0000606

Applying Multivariate Generalizability Theory to Psychological Assessments

Walter P. Vispoel[1], Hyeryung Lee[1], Hyeri Hong[2], and Tingting Chen[1]
[1] Department of Psychological and Quantitative Foundations, University of Iowa
[2] Department of Curriculum and Instruction, California State University Fresno

Abstract

Multivariate generalizability theory (GT) represents a comprehensive framework for quantifying score consistency, separating multiple sources contributing to measurement error, correcting correlation coefficients for such error, assessing subscale viability, and determining the best ways to change measurement procedures at different levels of score aggregation. Despite such desirable attributes, multivariate GT has rarely been applied when measuring psychological constructs and far less often than univariate techniques that are subsumed within that framework. Our purpose in this tutorial is to describe multivariate GT in a simple way and illustrate how it expands and complements univariate procedures. We begin with a review of univariate GT designs and illustrate how such designs serve as subcomponents of corresponding multivariate designs. Our empirical examples focus primarily on subscale and composite scores for objectively scored measures, but guidelines are provided for applying the same techniques to subjectively scored performance and clinical assessments. We also compare multivariate GT indices of score consistency and measurement error to those obtained using alternative GT-based procedures and across different software packages for analyzing multivariate GT designs. Our online supplemental materials include instruction, code, and output for common multivariate GT designs analyzed using *mGENOVA* and the *gtheory*, *glmmTMB*, *lavaan*, and related packages in R.

Translational Abstract

Accounting for multiple sources of measurement error is important when interpreting reliability and validity of results from measures of psychological constructs, and multivariate generalizability theory (GT) provides a comprehensive yet often neglected framework for facilitating such interpretations at multiple levels of score aggregation. We created this tutorial for readers who wish to enhance their understanding of multivariate GT techniques and apply them when developing, interpreting, and revising objectively or subjectively scored measures. Our examples illustrate how multivariate GT both subsumes and extends univariate GT by (a) simultaneously partitioning universe score and measurement error variance at subscale and composite levels, (b) defining global universes of generalization more precisely, (c) providing more appropriate indices of score consistency at composite levels, (d) correcting correlation coefficients for multiple sources of measurement error, (e) deriving confidence intervals for key parameters, (f) assessing subscale viability, (g) estimating how changes in measurement procedures might affect psychometric properties of scores, and (h) correcting for scale coarseness when such analyses are conducted within structural equation modeling frameworks. We also compare results for multivariate GT analyses to those using alternative GT techniques and across different computer packages for performing such analyses. Our online supplemental materials include instruction, code, and output from these packages for numerous crossed and nested designs relevant to psychological assessments.

Keywords: generalizability theory, multivariate analysis, psychometrics, Big Five Inventory, R programming

Supplemental materials: https://doi.org/10.1037/met0000606.supp

Over recent years, researchers have increasingly emphasized the importance of taking multiple sources of measurement error into account when interpreting reliability and validity of results from psychological assessments, and generalizability theory (GT; Brennan, 2001a; Cronbach et al., 1963, 1972; Shavelson & Webb, 1991) has served as a powerful and enduring framework for facilitating such interpretations using either objectively or subjectively scored measures. Applications of GT have typically focused on univariate

Walter P. Vispoel https://orcid.org/0000-0002-9415-251X

The authors thank the Iowa Measurement Research Foundation for providing research assistantship to support this project (Grant number: 520-14-2581-00000-88395100-5045-000-92045-20-0000).

The authors have no conflicts of interest to disclose. No work conducted for the study was preregistered.

Correspondence concerning this article should be addressed to Walter P. Vispoel, Department of Psychological and Quantitative Foundations, University of Iowa, 361 Lindquist Center, Iowa City, IA 52242-1529, United States. Email: walter-vispoel@uiowa.edu

analyses of individual scales that are appropriate when scores are considered individually or as part of a profile (see, e.g., Vispoel et al., 2018a, 2018b, 2018c, 2018d, 2019; Vispoel, Hong, et al., in press; Vispoel, Lee, Chen, & Hong, 2023a; Vispoel & Tao, 2013; Vispoel, Xu, & Kilinc, 2021; Vispoel, Xu, & Schneider, 2022a, 2022b). However, in situations in which scale scores are combined to form composites, multivariate GT analyses are preferred because they both subsume and extend univariate analyses in important ways. Our goal in this tutorial is to describe multivariate GT in a straightforward manner and demonstrate how it can be used effectively in research and practice.

Background

Overview

Brennan (2001a, p. 267) notes that Cronbach et al. (1972) considered development of multivariate GT as their most important contribution to the research literature, in part, because it both encompasses and expands univariate GT techniques. Although the overall computational framework for GT can be represented in analysis of variance/multivariate analysis of variance (ANOVA/MANOVA), linear mixed (aka, multilevel or hierarchical linear), or structural equation models, GT is applied differently than those models in practice. Univariate GT entails estimation of variance components to distinguish universe score and measurement error effects for individual measures, whereas multivariate GT includes estimation of both variance and covariance components to distinguish such effects at composite levels. Universe score variance in GT represents a single measure within univariate designs and multiple measures within multivariate designs. Each individual measure within a multivariate analysis is considered fixed with an underlying random-effects variance component design to represent facets contributing to measurement error. These random-effects univariate designs in turn are connected via covariance components to create the overall multivariate design.

GT differs from ANOVA/MANOVA, linear mixed, and structural equation modeling in that variance and covariance components are used primarily to derive indices of score consistency and dependability rather than to formally test hypotheses for the overall model or individual effects. Instead, confidence intervals can be derived to evaluate the trustworthiness of estimated variance components, covariance components, measurement error effects, and reliability coefficients. To set the stage for such applications, we begin with a brief overview of univariate GT designs and how they are subsequently incorporated into multivariate designs.

Univariate GT Designs

Basic Concepts

The purpose of univariate GT designs is to partition variance in observed scores into components representing explained and measurement error variance. Explained (non-error) variance is called *universe score* variance in GT and varies from *true score* variance in classical test theory in that it is intended to represent the entire domain(s) from which scores are sampled rather than the specific entities included in an investigation. Another key difference between the theories is that contributions to measurement error in GT are typically subdivided into multiple sources rather than considered strictly as a whole and together represent error in generalizing results to the broader assessment domains of interest.

GT analyses traditionally begin with a *generalizability* (or calibration) *study* in which universes of abmissable observations are defined and relevant variance component estimates are derived. Common universes of interest include tasks (e.g., items, parcels of items, split-halves, test forms), occasions, and raters. Each defined universe of generalization represents a facet that corresponds to a source of measurement error in the GT design. For objectively scored, self-report measures, facets typically include tasks and/or occasions. For subjectively scored measures (e.g., essay and other performance assessments), raters are often included as a facet along with tasks. Occasions sometimes serve as an additional facet in those designs when evaluation of intrarater consistency is also of interest. Once relevant variance components are derived, they can be used to estimate score consistency for both norm- and criterion-referencing purposes based on original or altered GT designs. These subsequent uses of GT results are often called *decision* (or application) *studies* (Cronbach et al., 1963).

Persons × Items Design

GT designs for self-report measures we discuss here consist of one or two measurement facets. Items represent the facet of interest in our first design, whereas items and occasions serve as facets in our second. Our single-facet design is formally referred to as a random-effects, *Persons* × *Items* ($p \times i$) design. Within this design, persons represent the *objects of measurement*, and items represent the single source of measurement error analyzed. The design is also fully crossed because all individuals answer all items.

Observed score variance for the $p \times i$ design can be partitioned at both item and item-mean score levels as shown in Equations 1 and 2. The letter I is capitalized in Equation 2 to signify averaging across item scores. Primes appear over n_s in these equations and elsewhere to indicate that they can be changed in decision studies. Within the original generalizability study, primes need not appear over n_s (Cronbach et al., 1963, p. 147):

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

$$p \times I \text{ design: item-mean score level: } \sigma^2_{Y_{pI}} = \sigma^2_p + \frac{\sigma^2_{pi,e}}{n'_i}, \quad (2)$$

where σ^2 = 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.

At the item level, observed score variance is an additive combination of three variance components (σ^2_p, $\sigma^2_{pi,e}$, σ^2_i) that, respectively, represent persons, overall contributions to measurement error, and items. The subscript for measurement error variance (*pi,e*) is intended to communicate that the error term reflects both interperson differences in item scores as well as sources of residual error not captured by other terms in the design. Partitioning at the item-mean score level excludes the variance component for items (σ^2_i) because item means are constants that do not affect relative differences in

scores across persons. Once obtained, variance components can be inserted into Equations 3–5 to derive three types of consistency indices: generalizability (G) coefficients, global dependability (D) coefficients, and cut-score-specific D coefficients:

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

$$\text{Global } D \text{ coefficient for } p \times I \text{ design} = \frac{\sigma_p^2}{\sigma_p^2 + \left(\frac{\sigma_{pi.e}^2}{n_i'} + \frac{\sigma_i^2}{n_i'}\right)}, \tag{4}$$

$$\text{Cut-score-specific } D \text{ coefficient for } p \times I \text{ design} = \frac{\sigma_p^2 + (\mu_Y - \text{cut score})^2}{\sigma_p^2 + (\mu_Y - \text{cut score})^2 + \left(\frac{\sigma_{pi.e}^2}{n_i'} + \frac{\sigma_i^2}{n_i'}\right)}. \tag{5}$$

In treatments of GT by other authors (see, e.g., Brennan, 2001a; Shavelson & Webb, 1991), G and D coefficients are often symbolized as $E\rho^2$ and Φ, respectively. These coefficients can range from 0 to 1, with higher values reflecting greater consistency. Values within parentheses for G and global D coefficients in Equations 3 and 4, respectively, represent *relative error* and *absolute error*.

The most suitable index of score consistency to report in practice would depend on the purpose of the assessment. G coefficients reflect relative differences in scores and are appropriate to report when scores are used for norm-referencing purposes (e.g., rank ordering). D coefficients, in contrast, are more informative for criterion-referencing purposes in which absolute levels of scores are relevant. Equations 3 and 4 illustrate that global D coefficients encompass both relative and absolute differences in scores and reduce to corresponding G coefficients when item means are equal (i.e., $\sigma_i^2 = 0$). With self-report measures within the present design, noticeable differences between G and global D coefficients would reflect nonequivalence of items in average levels of endorsement. When decisions are made on an individual person basis using targeted points along the score scale, cut-score-specific D coefficients would provide the most appropriate index of dependability for inferring whether an individual's universe score truly falls above or below the targeted cut-point. Such decisions are common when using cut scores for screening, selection, classification, or domain referencing purposes.

Persons × Items × Occasions Design

Although the three GT indices of score consistency described thus far are useful when data are limited to a single occasion, they do not reflect the major advantage of GT over classical test theory in distinguishing multiple sources of measurement error. To demonstrate such benefits, we now extend the $p \times i$ design by including occasions as an additional facet, thereby creating a random-effects, *Persons* × *Items* × *Occasions* ($p \times i \times o$) design. Once again, this design is fully crossed because each person completes all items on all occasions. Partitioning of variance for this design at item and item-mean score levels is shown in the following equations:

$$p \times i \times \text{o design: item score level:}$$
$$\sigma_{Y_{pio}}^2 = \sigma_p^2 + \sigma_{pi}^2 + \sigma_{po}^2 + \sigma_{pio,e}^2 + \sigma_i^2 + \sigma_o^2 + \sigma_{io}^2, \tag{6}$$

$$p \times I \times O \text{ design: item-mean score level:}$$
$$\sigma_{Y_{pIO}}^2 = \sigma_p^2 + \frac{\sigma_{pi}^2}{n_i'} + \frac{\sigma_{po}^2}{n_o'} + \frac{\sigma_{pio,e}^2}{n_i' n_o'}, \tag{7}$$

where $\sigma^2 =$ 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 letters I and O are capitalized in Equation 7 to signify averaging scores over both items and occasions. Note that the variance of scores at the item level in Equation 6 is now partitioned into seven components with one corresponding to persons (σ_p^2), three to interperson differences in item and/or occasion scores (σ_{pi}^2, σ_{po}^2, $\sigma_{pio,e}^2$), and three to absolute differences in item and/or occasion mean scores (σ_i^2, σ_o^2, σ_{io}^2). The subscript for $\sigma_{pio,e}^2$ emphasizes that it represents interperson differences in item and occasion scores as well as sources of residual error excluded from other terms in the design. At the item-mean level, partitioning again omits variance components for absolute differences in scores because means for items and occasions are constants after scores are aggregated that do not affect relative differences across persons. We provide formulas for G, global D, and cut-score-specific D coefficients for the $p \times I \times O$ design in Equations 8–10. Values within parentheses in Equations 8 and 9 again, respectively, represent relative and absolute error. As before, global D coefficients reduce to corresponding G coefficients when means for scores across facets do not differ (i.e., $\sigma_i^2 = \sigma_o^2 = \sigma_{io}^2 = 0$):

$$G \text{ coefficient for } p \times I \times O \text{ design} = \frac{\sigma_p^2}{\sigma_p^2 + \left(\frac{\sigma_{pi}^2}{n_i'} + \frac{\sigma_{po}^2}{n_o'} + \frac{\sigma_{pio,e}^2}{n_i' n_o'}\right)}, \tag{8}$$

$$\text{Global } D \text{ coefficient for } p \times I \times O \text{ design} = \frac{\sigma_p^2}{\sigma_p^2 + \left(\frac{\sigma_{pi}^2}{n_i'} + \frac{\sigma_{po}^2}{n_o'} + \frac{\sigma_{pio,e}^2}{n_i' n_o'} + \frac{\sigma_i^2}{n_i'} + \frac{\sigma_o^2}{n_o'} + \frac{\sigma_{io}^2}{n_i' n_o'}\right)}, \tag{9}$$

$$\text{Cut-score-specific } D \text{ coefficient for } p \times I \times O \text{ design} = \frac{\sigma_p^2 + (\mu_Y - \text{cut score})^2}{\sigma_p^2 + (\mu_Y - \text{cut score})^2 + \left(\frac{\sigma_{pi}^2}{n_i'} + \frac{\sigma_{po}^2}{n_o'} + \frac{\sigma_{pio,e}^2}{n_i' n_o'} + \frac{\sigma_i^2}{n_i'} + \frac{\sigma_o^2}{n_o'} + \frac{\sigma_{io}^2}{n_i' n_o'}\right)}. \tag{10}$$

A key distinction between G coefficients within $p \times i \times o$ and $p \times i$ designs for objectively scored measures (e.g., Likert-style questionnaires, multiple-choice tests) is that contributions to measurement

error are now subdivided into three categories, with σ^2_{pi}/n'_i representing specific-factor error, σ^2_{po}/n'_o representing transient error, and $\sigma^2_{pio,e}/(n'_i \times n'_o)$ representing random-response error (see, e.g., Le et al., 2009; Schmidt & Hunter, 1996; Schmidt et al., 2003). Specific-factor error represents enduring idiosyncratic characteristics of items unrelated to the overall construct(s) being measured. One example of such effects within objectively scored measures would be respondent-specific interpretations of words within items. Transient error stems from psychological and/or physical states (mood, feelings, motivation, fatigue, illness, etc.) that affect responses to all items within an occasion but not across occasions. Random-response error, in contrast, includes momentary fluctuations in cognitive efficiency, memory, attention, distractions, and other fleeting entities that influence a person's response to a given item at a specific time within an occasion. In measurement models such as latent state-trait theory, specific-factor and transient error, respectively, represent method and state effects that are usually treated as sources of explained/true-score variance (see, e.g., Vispoel, Xu, & Schneider, 2022a). When analyzing GT $p \times i \times o$ designs, specific-factor and/or transient error also can be treated as part of universe score variance if score consistency indices are appropriately adjusted.[1]

Further Estimation of Score Consistency and Proportions of Measurement Error

Once variance components are derived for the $p \times i \times o$ design, Equations 8–10 can be used in decision studies to estimate G and D coefficients for any desired combination of numbers of items (n'_i) and/or occasions (n'_o). These equations also can be altered to estimate proportions of relative observed score variance accounted for by individual sources of measurement error. This is accomplished by replacing person variance in the numerator of Equation 8 with the index for the targeted source of measurement error as shown in Equations 11–13. Similar substitutions can be made for estimating proportions of absolute error for D coefficients (see, e.g., Vispoel, Lee, Chen, & Hong, 2023b; Vispoel & Tao, 2013):

Proportion of specific-factor error for $p \times I \times O$ design

$$= \frac{\frac{\sigma^2_{pi}}{n'_i}}{\sigma^2_p + \left(\frac{\sigma^2_{pi}}{n'_i} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pio,e}}{n'_i n'_o}\right)}, \tag{11}$$

Proportion of transient error for $p \times I \times O$ design

$$= \frac{\frac{\sigma^2_{po}}{n'_o}}{\sigma^2_p + \left(\frac{\sigma^2_{pi}}{n'_i} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pio,e}}{n'_i n'_o}\right)}, \tag{12}$$

Proportion of random-response error for $p \times I \times O$ design

$$= \frac{\frac{\sigma^2_{pio,e}}{n'_i n'_o}}{\sigma^2_p + \left(\frac{\sigma^2_{pi}}{n'_i} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pio,e}}{n'_i n'_o}\right)}. \tag{13}$$

Restricting Universes of Generalization

Another common application of GT $p \times i \times o$ designs in decision studies is to estimate score consistency for more restricted universes of generalization. Indices described up to this point reflect generalization of results across both items and occasions but can be altered to restrict generalization to just items or just occasions. This is easily accomplished by merging relative error variance components for omitted measurement facets with person variance in the numerator and denominator of equations for G and D coefficients and limiting subsequent error variance components in the denominators of the coefficients to only those for retained facets. We illustrate alterations to G and global D coefficients for generalizing across just items and just occasions in Equations 14–17. Similar adjustments can be made to the formulas for cut-score-specific D coefficients:

$$G \text{ coefficient (items only)} = \frac{\sigma^2_p + \frac{\sigma^2_{po}}{n'_o}}{\sigma^2_p + \frac{\sigma^2_{po}}{n'_o} + \left(\frac{\sigma^2_{pi}}{n'_i} + \frac{\sigma^2_{pio,e}}{n'_i n'_o}\right)}, \tag{14}$$

Global D coefficient (items only)

$$= \frac{\sigma^2_p + \frac{\sigma^2_{po}}{n'_o}}{\sigma^2_p + \frac{\sigma^2_{po}}{n'_o} + \left(\frac{\sigma^2_{pi}}{n'_i} + \frac{\sigma^2_{pio,e}}{n'_i n'_o} + \frac{\sigma^2_{i}}{n'_i} + \frac{\sigma^2_{io}}{n'_i n'_o}\right)}, \tag{15}$$

G coefficient (occasions only)

$$= \frac{\sigma^2_p + \frac{\sigma^2_{pi}}{n'_i}}{\sigma^2_p + \frac{\sigma^2_{pi}}{n'_i} + \left(\frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pio,e}}{n'_i n'_o}\right)}, \tag{16}$$

Global D coefficient (occasions only)

$$= \frac{\sigma^2_p + \frac{\sigma^2_{pi}}{n'_i}}{\sigma^2_p + \frac{\sigma^2_{pi}}{n'_i} + \left(\frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pio,e}}{n'_i n'_o} + \frac{\sigma^2_{o}}{n'_o} + \frac{\sigma^2_{io}}{n'_i n'_o}\right)}. \tag{17}$$

Terms are arranged in Equations 14–17 to illustrate that person and transient error variances are confounded as are specific-factor

[1] Within latent state-trait theory orthogonal method models involving items and occasions, score consistency coefficients that treat trait and method effects as explained (i.e., non-error) variance are referred to as *total consistency* coefficients, those that treat trait and states effects as explained variance as *common reliability* coefficients, and those that treat trait, method, and state effects as explained variance as *reliability* coefficients (Geiser & Lockhart, 2012; Steyer et al., 1992, 2012). At the total score level within GT $p \times I \times O$ designs, latent state-trait theory total consistency coefficients would be analogous to G coefficients in which results are generalized across items but not occasions, and common reliability coefficients would correspond to G coefficients in which results are generalized across occasions but not items. Vispoel, Xu, & Schneider, 2022a discuss these and other relationships between GT and latent state-trait theory in detail.

and random-response error variances within score consistency indices when based solely on items, whereas person and specific-factor error variances are confounded as are transient and random-response error variances when those indices are based solely on occasions. These same issues apply to conventional single-occasion (alpha, omega, and split-half) and test–retest reliability estimates. Occasion effects are hidden within single-occasion indices, and item effects are hidden within test–retest coefficients (Brennan, 2001a; Vispoel et al., 2018a).[2]

Multivariate GT Designs

Basic Concepts

Multivariate GT designs allow for simultaneous derivation of *G* coefficients, *D* coefficients, and proportions of measurement error at both subscale and overall composite score levels. Subscales within the domain encompassed by the composite score essentially serve as fixed strata that help define the overall domain of interest more precisely, while still providing the same information as univariate GT analyses for each individual subscale. Such analyses are appropriate whenever results are interpreted at both profile and combined score levels. In addition to applications within decision studies for univariate analyses already discussed, multivariate GT reveals how each subscale is weighted when computing composite score variances, allows for variations in such weights, and provides estimates of correlations between universe scores for pairs of subscales corrected for all relevant sources of measurement error.

Variance and Covariance Component Matrices

Analysis of composite scores in multivariate GT requires estimation of variance components for each subscale and covariance components between subscales. Each variance component for a univariate analysis is replaced with a variance–covariance matrix in the multivariate analysis. There would be a separate matrix (Σ) for Σ_p, $\Sigma_{pi,e}$, and Σ_i in the $p \times i$ design, and for Σ_p, Σ_{pi}, Σ_{po}, $\Sigma_{pio,e}$, Σ_i, Σ_o, and Σ_{io} in the $p \times i \times o$ design. If we had three subscales, each matrix would be in the form shown in the following equation:

$$\text{Variance-covariance matrix for three subscales} = \begin{pmatrix} \sigma^2_{S_1} & \sigma_{S_1S_2} & \sigma_{S_1S_3} \\ \sigma_{S_2S_1} & \sigma^2_{S_2} & \sigma_{S_2S_3} \\ \sigma_{S_3S_1} & \sigma_{S_3S_2} & \sigma^2_{S_3} \end{pmatrix}. \tag{18}$$

The diagonal elements in the matrices are the same variance components that would be derived from a univariate analysis for each subscale, whereas the off-diagonal elements represent covariances between subscales. Composite score variances from any one of these matrices would represent a weighted sum of all relevant subscale score variances and covariances appearing in the matrix. Subscale variances would be multiplied by their squared weights and covariances by the product of weights for each pair of subscales represented. Because covariances are repeated in the upper and lower triangles of the matrix, each weighted covariance can be multiplied by two when computing the variance for the composite. As a result, the variance for the composite would equal the sum of the variances for each subscale multiplied by their squared weights plus two times the sum of each possible covariance times the product of weights for the pair of subscales included (see Equation 19). Due to the redundancy of terms in the lower and upper triangles of matrices, those in upper triangle are often omitted when presenting matrices for multivariate GT analyses (see Table 2 appearing later in this article).

$$\begin{aligned} \sigma^2_{\text{composite}} &= \sum_{i=1}^{n_S} w^2_{S_i}\sigma^2_{S_i} + \sum_{i=1}^{n_S}\sum_{j\neq i}^{n_S} w_{S_i}w_{S_j}\sigma_{S_i,S_j} \\ &= \sum_{i=1}^{n_S} w^2_{S_i}\sigma^2_{S_i} + 2\left(\sum_{i=1}^{n_S}\sum_{j>i}^{n_S} w_{S_i}w_{S_j}\sigma_{S_i,S_j}\right), \end{aligned} \tag{19}$$

where S = subscale, w = weight, and n_s = number of subscales.

The computations shown in Equation 19 follow standard rules for determining the variance of a composite score when subscale scores are weighted and combined (see, e.g., Allen & Yen, 2002; Crocker & Algina, 1986; Gulliksen, 1950; Lord et al., 1968; Nunnally & Bernstein, 1994). If we applied this formula when computing a variance for composite scores based on three subscale scores, Equation 19 could be expressed in the long form shown below:

$$\begin{aligned} \sigma^2_{\text{composite}} = {} & w^2_{S_1}\sigma^2_{S_1} + w^2_{S_2}\sigma^2_{S_2} + w^2_{S_3}\sigma^2_{S_3} \\ & + 2(w_{S_1}w_{S_2}\sigma_{S_1S_2} + w_{S_1}w_{S_3}\sigma_{S_1S_3} + w_{S_2}w_{S_3}\sigma_{S_2S_3}). \end{aligned} \tag{20}$$

This same procedure would be applied for each variance–covariance matrix in a multivariate GT design to derive the corresponding variance component for the composite score.

Score Consistency and Measurement Error Indices for Composites

When designs are balanced with each subscale (S) having the same number of items (i), variance components for composite scores can be substituted into Equations 3–5 and 8–13 from the embedded univariate designs to determine score consistency indices and relevant proportions of measurement error for the composite score (C) as shown in Equations 21–26[3]:

$$\begin{aligned} &\text{Composite } G \text{ coefficient in the multivariate } \textit{persons} \\ &\quad \times \textit{Items} \times \textit{Occasions} \text{ design} \\ &= \frac{\sigma^2_{C(p)}}{\sigma^2_{C(p)} + \left(\dfrac{\sigma^2_{C(pi)}}{n'_{S(i)}} + \dfrac{\sigma^2_{C(po)}}{n'_{S(o)}} + \dfrac{\sigma^2_{C(pio,e)}}{n'_{S(i)}n'_{S(o)}}\right)}. \end{aligned} \tag{21}$$

[2] Conventional and GT reliability estimates that take just task, just occasion, and both task and occasion effects into account are sometimes respectively referred to as coefficients of equivalence (CEs), stability (CSs), and equivalence and stability (CESs). CEs and CESs in classical test theory would typically represent *classically* equivalent or parallel measures, whereas those in GT would usually represent *randomly* equivalent measures.

[3] With unbalanced designs in which the number of items varies with subscales, the value for $n'_{S(i)}$ in the equations shown in this tutorial is replaced with a *harmonic mean* that equals the number of subscales divided by the sum of the reciprocals of the number of items within each scale. For example, if there are three subscales with three, four, and five items, the harmonic mean would equal 3/(1/3 + 1/4 + 1/5) or 3.83.

Composite global D coefficient in the multivariate *persons* × *Items* × *Occasions* design

$$= \frac{\sigma^2_{C(p)}}{\sigma^2_{C(p)} + \left(\frac{\sigma^2_{C(pi)}}{n'_{S(i)}} + \frac{\sigma^2_{C(po)}}{n'_{S(o)}} + \frac{\sigma^2_{C(pio,e)}}{n'_{S(i)}n'_{S(o)}} + \frac{\sigma^2_{C(i)}}{n'_{S(i)}} + \frac{\sigma^2_{C(o)}}{n'_{S(o)}} + \frac{\sigma^2_{C(io)}}{n'_{S(i)}n'_{S(o)}}\right)}. \quad (22)$$

General equation for global D coefficients in univariate and multivariate GT designs:

$$\text{Global } D \text{ coefficient} = \frac{\text{Universe score variance}}{\text{Universe score variance + absolute error variance}}. \quad (28)$$

Composite cut-score-specific D coefficient in the multivariate *persons* × *Items* × *Occasions* design

$$= \frac{\sigma^2_{C(p)} + (\mu_Y - \text{cut score})^2}{\sigma^2_{C(p)} + (\mu_Y - \text{cut score})^2 + \left(\frac{\sigma^2_{C(pi)}}{n'_{S(i)}} + \frac{\sigma^2_{C(po)}}{n'_{S(o)}} + \frac{\sigma^2_{C(pio,e)}}{n'_{S(i)}n'_{S(o)}} + \frac{\sigma^2_{C(i)}}{n'_{S(i)}} + \frac{\sigma^2_{C(o)}}{n'_{S(o)}} + \frac{\sigma^2_{C(io)}}{n'_{S(i)}n'_{S(o)}}\right)}. \quad (23)$$

Composite proportion of specific-factor error in the multivariate *persons* × *Items* × *Occasions* design

$$= \frac{\frac{\sigma^2_{C(pi)}}{n'_{S(i)}}}{\sigma^2_{C(p)} + \left(\frac{\sigma^2_{C(pi)}}{n'_{S(i)}} + \frac{\sigma^2_{C(po)}}{n'_{S(o)}} + \frac{\sigma^2_{C(pio,e)}}{n'_{S(i)}n'_{S(o)}}\right)}. \quad (24)$$

Composite proportion of transient error in the multivariate *persons* × *Items* × *Occasions* design

$$= \frac{\frac{\sigma^2_{C(po)}}{n'_{S(o)}}}{\sigma^2_{C(p)} + \left(\frac{\sigma^2_{C(pi)}}{n'_{S(i)}} + \frac{\sigma^2_{C(po)}}{n'_{S(o)}} + \frac{\sigma^2_{C(pio,e)}}{n'_{S(i)}n'_{S(o)}}\right)}. \quad (25)$$

Composite proportion of random-response error in the multivariate *persons* × *Items* × *Occasions* design

$$= \frac{\frac{\sigma^2_{C(pio,e)}}{n'_{S(i)}n'_{S(o)}}}{\sigma^2_{C(p)} + \left(\frac{\sigma^2_{C(pi)}}{n'_{S(i)}} + \frac{\sigma^2_{C(po)}}{n'_{S(o)}} + \frac{\sigma^2_{C(pio,e)}}{n'_{S(i)}n'_{S(o)}}\right)}. \quad (26)$$

As before, values within parentheses for the G coefficient in Equation 21 represent relative error, whereas those for global and cut-score-specific D coefficients in Equations 22 and 23 represent absolute error. General formulas for these coefficients in both univariate or multivariate designs are shown in Equations 27 and 28.

General equation for G coefficients in univariate and multivariate GT designs:

$$G \text{ coefficient} = \frac{\text{Universe score variance}}{\text{Universe score variance + relative error variance}}. \quad (27)$$

Disattenuated Correlation Coefficients

The person matrix within multivariate GT designs also can be used to derive correlation coefficients between subscale scores corrected for relevant sources of measurement error by dividing their covariance component by the square root of the product of their variance components. For example, if Equation 18 represented a person matrix, the disattenuated correlation between Subscales 1 and 2 would equal $\sigma_{S_1S_2}/\sqrt{\sigma^2_{S_1} \times \sigma^2_{S_2}}$.[4]

Additional Notation

Formal notation for multivariate GT designs is extended to indicate whether persons and measurement facets (items and occasions here) are crossed with or nested under subscales in the design. Closed circles are used when persons or facets are crossed with subscales, whereas open circles are used when they are nested. In the present situation in which persons and occasions are crossed with subscales, and items are nested within subscales, multivariate designs corresponding to the previous univariate designs would be labeled $p^\bullet \times i^\circ$ and $p^\bullet \times i^\circ \times o^\bullet$, respectively. When a measurement facet is nested within subscales, as is the case with items here, covariances would be excluded within any matrix related to that measurement facet (see Table 2 and Figure 3 appearing later in this article). This would not be the case in a situation in which the same item stems are used for all subscales. In such instances, the previous designs would, respectively, be labeled $p^\bullet \times i^\bullet$ and $p^\bullet \times i^\bullet \times o^\bullet$. Similarly, when the same raters use the same multidimensional rubric to score a series of performance tasks for the same group of individuals, the multivariate design would be labeled $p^\bullet \times r^\bullet \times t^\bullet$, where r represents raters and t represents tasks. Within these completely crossed designs, both variances

[4] When measurement facets are crossed with subscales, correlations also can be derived between corresponding measurement error terms. For example, this might be of interest when occasions are crossed with subscales to determine the extent to which transient error/state effects are consistent across subscales (see, e.g., Vispoel et al., 2018c).

Table 1
Descriptive Statistics and Conventional Reliability Estimates for BFI-2 Composite and Subscale Scores

		Index			
Occasion/scale	Number of items	*M*: scale (item)	*SD*: scale (item)	Alpha	Test–retest
Time 1					
Extraversion	12	39.80 (3.32)	8.75 (0.73)	0.86	0.90
Assertiveness	4	12.77 (3.19)	3.45 (0.86)	0.77	0.84
Energy Level	4	14.35 (3.59)	3.13 (0.78)	0.67	0.85
Sociability	4	12.68 (3.17)	4.03 (1.01)	0.80	0.90
Time 2					
Extraversion	12	40.01 (3.33)	8.65 (0.72)	0.87	
Assertiveness	4	12.82 (3.20)	3.43 (0.86)	0.77	
Energy Level	4	14.38 (3.59)	3.18 (0.80)	0.73	
Sociability	4	12.82 (3.20)	3.88 (0.97)	0.82	

Note. BFI-2 = Big Five Inventory-2.

and covariances would be estimated within each matrix. If different subscales with different items are administered on different occasions or different raters evaluate different tasks, open circles would be used to represent the facets (i.e., $p^\bullet \times i^\circ \times o^\circ$ and $p^\bullet \times r^\circ \times t^\circ$). To clarify multivariate GT designs further, Brennan (2001a, p. 274) provided added notation for their overall univariate counterparts using the letter v to represent subscales as a fixed measurement facet. With such notation, univariate designs corresponding to multivariate $p^\bullet \times i^\circ$, $p^\bullet \times i^\bullet$, $p^\bullet \times i^\circ \times o^\bullet$, $p^\bullet \times i^\bullet \times o^\bullet$, and $p^\bullet \times i^\circ \times o^\circ$ designs would be labeled, respectively, as: $p \times (i{:}v)$, $p \times i \times v$, $p \times (i{:}v) \times o$, $p \times i \times o \times v$, and $p \times [(i \times o){:}v]$.

Actual Weighting of Subscale Scores

Multivariate GT provides great flexibility in assigning weights to subscale scores based on previous empirical data or expert judgments catered to maximizing either the reliability or validity of those scores (see, e.g., Baldwin, 2015). Without such directives, initial weights for subscales when computing *G* and *D* coefficients are usually based on the proportion of items for each subscale represented in the composite score. For example, if each subscale has the same number of items, and there are three subscales, the initial weight for each subscale would equal 1/3. However, the actual weighting for a subscale within the final composite would also

Table 2
Variance–Covariance Matrices for $p^\bullet \times i^\circ$ and $p^\bullet \times i^\circ \times o^\bullet$ Designs

		Subscale within $p^\bullet \times i^\circ$ design			Subscale within $p^\bullet \times i^\circ \times o^\bullet$ design		
Matrix	Subscale	Assertiveness	Energy level	Sociability	Assertiveness	Energy level	Sociability
$\hat{\Sigma}_p$	Assertiveness	.572 (.572)			.531 (.531)		
	Energy level	.289 (.289)	.411 (.411)		.263 (.263)	.421 (.421)	
	Sociability	.503 (.503)	.414 (.414)	.808 (.808)	.468 (.468)	.392 (.392)	.754 (.754)
$\hat{\Sigma}_{pi,e}$ or $\hat{\Sigma}_{pi}$	Assertiveness	.694 (.173)			.366 (.091)		
	Energy level	—	.797 (.199)		—	.427 (.107)	
	Sociability	—	—	.824 (.206)	—	—	.505 (.126)
$\hat{\Sigma}_{po}$	Assertiveness				.038 (.038)		
	Energy level				.020 (.020)	.015 (.015)	
	Sociability				.023 (.023)	.026 (.026)	.036 (.036)
$\hat{\Sigma}_{pio,e}$	Assertiveness				.323 (.081)		
	Energy level				—	.315 (.079)	
	Sociability				—	—	.249 (.062)
$\hat{\Sigma}_i$	Assertiveness	.037 (.009)			.036 (.009)		
	Energy level	—	.100 (.025)		—	.094 (.023)	
	Sociability	—	—	.218 (.054)	—	—	.221 (.055)
$\hat{\Sigma}_o$	Assertiveness				.000 (.000)		
	Energy level				.000 (.000)	.000 (.000)	
	Sociability				.000 (.000)	.000 (.000)	.000 (.000)
$\hat{\Sigma}_{io}$	Assertiveness				−.001 (.000)		
	Energy level				—	.001 (.000)	
	Sociability				—	—	−.001 (.000)

Note. Covariance components are only estimated when measurement facets are crossed with subscales. In the present multivariate GT designs, persons and occasions are crossed with subscales, but items are nested. As a result, covariance components are excluded in the matrices that include items as a facet. Values outside the parentheses are the original variance or covariance components derived from *mGENOVA*. Those within parentheses are ones for a decision study in which relevant components are divided by number of items (four), number of occasions (one), or the product of those values.

depend on the values of variances and covariances for subscale scores. In multivariate GT, these operational weights are often referred to as *effective weights* (*ew*; Brennan, 2001a, pp. 306–307). In essence, these weights represent the real proportionate contribution of a subscale to a targeted composite index of variance. Such weights are typically derived for universe score, relative error, and absolute error and are provided as standard output when using the *mGENOVA* multivariate GT package (Brennan, 2001b).

Computing an effective weight would entail extracting and combining variance and covariance terms for a given subscale within a composite variance formula and dividing that sum by the corresponding composite variance. For example, suppose we wanted to determine the contribution of the first subscale to a composite variance based on three subscales. Using the composite score variance defined in Equation 20, the effective weight (*ew*) for Subscale 1 (S_1) could be derived using the following equation:

$$ew \text{ for } S_1 = \frac{w_{S_1}^2\sigma_{S_1}^2 + w_{S_1}w_{S_2}\sigma_{S_1S_2} + w_{S_1}w_{S_3}\sigma_{S_1S_3}}{w_{S_1}^2\sigma_{S_1}^2 + w_{S_2}^2\sigma_{S_2}^2 + w_{S_3}^2\sigma_{S_3}^2 + 2(w_{S_1}w_{S_2}\sigma_{S_1S_2} + w_{S_1}w_{S_3}\sigma_{S_1S_3} + w_{S_2}w_{S_3}\sigma_{S_2S_3})}. \quad (29)$$

Note that the numerator includes the variance for Subscale 1 and its covariances with Subscales 2 and 3. A parallel approach would be used to derive effective weights for the remaining subscales. This process would only include the person matrix when determining effective weights for universe scores but would need to be repeated using the matrix for every relevant variance term and then combining them to determine effective weights for relative and absolute error.

In Equation 30, we provide a general formula for how to compute effective weights for universe score variance from the p matrix. This same equation applies to $p^\bullet \times I^\circ$, $p^\bullet \times I^\circ \times O^\bullet$ and other crossed and nested multivariate designs because computations are limited to a single matrix. Equation 31 represents an example of using the formula to determine the effective weight for Subscale 1 when three subscales are included in the design. The same procedure would be repeated to determine the effective weight for each subscale.

Universe score effective weight formulas for crossed persons multivariate GT designs

General form:

$$ew_{s_i}(p) = \frac{w_{s_i}^2\sigma_{s_i}^2(p) + w_{s_i}\sum_{s_j=1,\ s_j\neq s_i}^{s_n} w_{s_j}\sigma_{s_is_j}(p)}{\sigma_c^2(p)}. \quad (30)$$

Example using Subscale 1:

$$ew_{s_1}(p) = \frac{w_{s_1}^2\sigma_{s_1}^2(p) + w_{s_1}w_{s_2}\sigma_{s_1s_2}(p) + w_{s_1}w_{s_3}\sigma_{s_1s_3}(p)}{\sigma_c^2(p)} = \frac{w_{s_1}[w_{s1}\sigma_{s_1}^2(p) + w_{s_2}\sigma_{s_1s_2}(p) + w_{s_3}\sigma_{s_1s_3}(p)]}{\sigma_c^2(p)}. \quad (31)$$

Formulas for deriving effective weights for absolute error are more complicated than those for relative error because terms are included for both relative and absolute differences in scores. These formulas also become increasingly more complex as additional facets are added to a design. We provide a general formula for determining effective weights for relative error in the $p^\bullet \times I^\circ \times O^\bullet$ design with three subscales in Equation 32 and apply that formula to Subscale 1 in Equation 33. Complete sets of equations for determining effective weights for universe score, relative error, and absolute error for one- and two-facet crossed and nested multivariate GT designs are provided in our online supplemental materials.

Relative error effective weight formulas for the $p^\bullet \times I^\circ \times O^\bullet$ multivariate GT design

General form:

$$ew_{s_i}(\text{relative error}) = \frac{\frac{w_{s_i}^2\sigma_{s_i}^2(pi)}{n'_{s_i(i)}} + \frac{w_{s_i}\left[w_{s_i}\sigma_{s_i}^2(po) + \sum_{s_j=1,s_j\neq s_i}^{s_n} w_{s_j}\sigma_{s_is_j}(po)\right]}{n'_{s_i(o)}} + \frac{w_{s_i}^2\sigma_{s_i}^2(pio,e)}{n'_{s_i(i)}n'_{s_i(o)}}}{\frac{\sigma_c^2(pi)}{n'_{s(i)}} + \frac{\sigma_c^2(po)}{n'_{s(o)}} + \frac{\sigma_c^2(pio,e)}{n'_{s(i)}n'_{s(o)}}}. \quad (32)$$

Example using Subscale 1:

$$ew_{s_1}(\text{relative error}) = \frac{\frac{w_{s_1}^2\sigma_{s_1}^2(pi)}{n'_{s_1(i)}} + \frac{w_{s_1}^2\sigma_{s_1}^2(po) + w_{s_1}w_{s_2}\sigma_{s_1s_2}(po) + w_{s_1}w_{s_3}\sigma_{s_1s_3}(po)}{n'_{s_1(o)}} + \frac{w_{s_1}^2\sigma_{s_1}^2(pio,e)}{n'_{s_1(i)}n'_{s_1(o)}}}{\frac{\sigma_c^2(pi)}{n'_{s(i)}} + \frac{\sigma_c^2(po)}{n'_{s(o)}} + \frac{\sigma_c^2(pio,e)}{n'_{s(i)}n'_{s(o)}}}. \quad (33)$$

Changing Facet Conditions and Restricting Universes of Generalization

Calculating composite score *G* and *D* coefficients and proportions of measurement error for changes made to numbers of facet conditions is accomplished in the same manner as in univariate designs by specifying desired numbers of items and/or occasions in relevant equations. Deriving composite score *G* and global *D* coefficients for restricted universes again involves merging relative error variance components for omitted measurement facets with person variance in the numerator and denominator of those coefficients and limiting subsequent error variance components to only those for retained facets in their denominators as shown in Equations 34–37 for balanced designs (see Footnote 3 for alterations in unbalanced designs).

Patterns of confounding among person and measurement error variances for composite scores match those described earlier for corresponding individual subscale scores.

Composite G coefficient (items only)

$$= \frac{\sigma^2_{C(p)} + \frac{\sigma^2_{C(po)}}{n'_{S(o)}}}{\sigma^2_{C(p)} + \frac{\sigma^2_{C(po)}}{n'_{S(o)}} + \left(\frac{\sigma^2_{C(pi)}}{n'_{S(i)}} + \frac{\sigma^2_{C(pio,e)}}{n'_{S(i)}n'_{S(o)}}\right)}, \quad (34)$$

Composite global D coefficient (items only)

$$= \frac{\sigma^2_{C(p)} + \frac{\sigma^2_{C(po)}}{n'_{S(o)}}}{\sigma^2_{C(p)} + \frac{\sigma^2_{C(po)}}{n'_{S(o)}} + \left(\frac{\sigma^2_{C(pi)}}{n'_{S(i)}} + \frac{\sigma^2_{C(pio,e)}}{n'_{S(i)}n'_{S(o)}} + \frac{\sigma^2_{C(i)}}{n'_{S(i)}} + \frac{\sigma^2_{C(io)}}{n'_{S(i)}n'_{S(o)}}\right)}, \quad (35)$$

Composite G coefficient (occasions only)

$$= \frac{\sigma^2_{C(p)} + \frac{\sigma^2_{C(pi)}}{n'_{S(i)}}}{\sigma^2_{C(p)} + \frac{\sigma^2_{C(pi)}}{n'_{S(i)}} + \left(\frac{\sigma^2_{C(po)}}{n'_{S(o)}} + \frac{\sigma^2_{C(pio,e)}}{n'_{S(i)}n'_{S(o)}}\right)}, \quad (36)$$

Composite global D coefficient (occasions only)

$$= \frac{\sigma^2_{C(p)} + \frac{\sigma^2_{C(pi)}}{n'_{S(i)}}}{\sigma^2_{C(p)} + \frac{\sigma^2_{C(pi)}}{n'_{S(i)}} + \left(\frac{\sigma^2_{C(po)}}{n'_{S(o)}} + \frac{\sigma^2_{C(pio,e)}}{n'_{S(i)}n'_{S(o)}} + \frac{\sigma^2_{C(o)}}{n'_{S(o)}} + \frac{\sigma^2_{C(io)}}{n'_{S(i)}n'_{S(o)}}\right)}. \quad (37)$$

Subscale Viability

An important issue to consider when using both composite and subscale scores in practice is whether subscale scores provide useful information beyond that provided by the composite score. For subscale scores to be useful, they should possess good psychometric qualities and be distinct from one another. Verification of these attributes can take many forms including conventional procedures for deriving evidence of reliability and validity for each scale (American Educational Research Association, et al., 2014); factor analytic and related techniques to examine score dimensionality (Rodriguez et al., 2016a, 2016b); evaluation of the diagnostic effectiveness of subscale score profiles (Brennan, 2001a; Jiang & Raymond, 2018; Raymond & Jiang, 2020); comparisons of the concurrent/predictive validity for composite scores versus individual and/or all subscale scores considered collectively (Soto & John, 2017); and procedures suggested by Haberman (2008) (also see Haberman & Sinharay, 2010; Sinharay, 2019) to determine whether observed subscale scores better estimate a subscale's true scores than do observed composite scores. Because these techniques can lead to different conclusions regarding scale viability (see, e.g., Jiang & Raymond, 2018; Raymond & Jiang, 2020; Rodriguez et al., 2016a; Vispoel et al., 2022), the purpose of the assessment and nature of reported indices should be carefully considered.

We illustrate Haberman's (2008) procedure here because it is based on classical test theory and can be readily extended to GT designs by substituting universe score for true score estimation. The procedure simply requires sample variances, covariances, and reliability coefficients. Haberman's examples were limited to single occasions and KR-20 reliability estimates (Kuder & Richardson, 1937), but the same principles can be applied to other conventional reliability estimates (e.g., alpha, test–retest, parallel-form, and split-halves) and G coefficients that take multiple sources of measurement error into account.

Haberman's (2008) procedure involves computation and comparison of indices for a subscale and its related composite scores that represent proportional reduction in mean-squared error (PRMSE) when estimating the subscale's true scores. Subscale viability is supported when the PRMSE value (i.e., reduction in error) for the subscale exceeds that for the composite scale. The PRMSE value for a subscale score reduces to its corresponding reliability coefficient (conventional or GT-based), and the PRMSE value for the composite score can be estimated using the general formula shown in the following equation:

$$\text{PRMSE}_{(\text{Composite})} = \frac{\left(\hat{\sigma}^2_{(\text{Subscale}_i)} \times \text{Estimated Reliability}_{(\text{Subscale}_i)} + \sum_{j \neq i} \hat{\sigma}_{(\text{Subscale}_i,\text{Subscale}_j)}\right)^2}{\hat{\sigma}^2_{(\text{Subscale}_i)} \times \text{Estimated Reliability}_{(\text{Composite})} \times \hat{\sigma}^2_{(\text{Composite})}}. \quad (38)$$

To gauge the extent to which a subscale's observed scores might better represent the subscale's true scores, Feinberg and Wainer (2014) proposed forming a complementary *value-added ratio* (VAR) by dividing the subscale PRMSE by the composite PRMSE (see Equation 39). Subscale scores are considered increasingly useful as VARs deviate upwardly from 1.00:

$$\text{Value-added ratio (VAR)} = \frac{\text{PRMSE(subscale)}}{\text{PRMSE(composite)}}. \quad (39)$$

Empirical Examples of Multivariate GT Analyses

Data Source and Measures

Illustrations to follow are based on data collected from 346 college students (76% female; 75% Caucasian; $M_{age} = 21.15$) who completed online versions of the recently expanded form of the Big Five Inventory-2 (BFI-2; Soto & John, 2017) on two occasions a week apart. The study was approved by the governing Institutional Review Board, and all participants provided informed consent before responding to the measures. The study was not preregistered, and inquiries about accessibility to the data should be forwarded to Walter P. Vispoel. For sake of brevity, we confine our analyses here to the BFI-2's Extraversion composite scale and its three nested subscales: Assertiveness, Energy Level, and Sociability. Parallel analyses could be conducted for any of the remaining personality domains included in the BFI-2 or for composite and nested subscale scores from any other instrument (see, e.g., Vispoel, Lee, Xu, & Hong 2023; Vispoel, Lee, & Hong, in press).

The BFI-2's Extraversion composite scale has 12 items, and each subscale has four items. To control for possible acquiescence bias, items from all scales are equally balanced for positive and negative phasing. Descriptive statistics (means and standard deviations) and conventional reliability estimates (alpha, test–retest) for these scales are provided in Table 1. Overall, these indices are very much in line with those reported for college students by Soto and John (2017) and by other researchers in more recent investigations (see, e.g., Vispoel, Lee, Xu, & Hong, 2023; Vispoel, Lee, & Hong, in press; Vispoel, Lee, Chen, & Hong, 2023c; Vispoel, Xu, & Schneider, 2022b).

Main Analyses and Results

We focus exclusively on results from multivariate GT analyses here because univariate analyses for individual subscales are embedded within those analyses. We conducted initial analyses for $p^{\bullet} \times i^{\circ}$ and $p^{\bullet} \times i^{\circ} \times o^{\bullet}$ designs using the *mGENOVA* package (Brennan, 2001b) and provide code and output for those analyses in our online supplemental materials. Estimates of variance and covariance components from the analyses are provided in Table 2. Note that a couple of very small negative variances (−.001) appear in the $\hat{\Sigma}_{io}$ matrix for the $p^{\bullet} \times i^{\circ} \times o^{\bullet}$ design. Common ways to handle such components in GT analyses are to set them equal to zero or retain the negative values when computing score consistency indices in decision studies (Brennan, 2001a; Shavelson et al., 1992). An alternative way to avoid negative variance components altogether is to use restricted maximum likelihood (Marcoulides, 1990; Vispoel, Xu, & Schneider, 2022b) or Bayesian procedures when estimating the components (see, e.g., Jiang & Skorupski, 2018; LoPilato et al., 2015; Vispoel et al., 2022; and our online supplemental materials for examples of code for running such analyses). Because the negative variance components were so low in magnitude here, results for *G* and *D* coefficients were essentially the same when retaining or setting them to zero. Computation of relevant composite level variance components, *G* coefficients, global *D* coefficients, and proportions of measurement error for the $p^{\bullet} \times I^{\circ}$ and $p^{\bullet} \times I^{\circ} \times O^{\bullet}$ designs is illustrated in Table 3. We provide additional formulas for these indices within $p^{\bullet} \times i^{\bullet}$, $p^{\bullet} \times i^{\bullet} \times o^{\bullet}$, and $p^{\bullet} \times i^{\circ} \times o^{\circ}$ designs in our online supplemental materials.

Multivariate GT Analyses for the $p^{\bullet} \times i^{\circ}$ Design

As noted earlier, multivariate GT analyses for the $p^{\bullet} \times i^{\circ}$ design entail estimation of the following variance–covariance matrices: Σ_p, $\Sigma_{pi,e}$, and Σ_i. Elements in the diagonal of those matrices represent variance components for separate univariate GT analyses for each subscale that allow for calculation of *G*, global *D*, and cut-score-specific *D* coefficients. Combining variance and covariance components from the same matrices permits calculation of corresponding coefficients for the composite score. We provide *G* and global *D* coefficients for subscale scores with four items and composite scores with 12 items in Table 4, and associated cut-score-specific *D* coefficients expressed on the item scale metric in the first graph in Figure 1. The same coefficients in the figure can be referenced to the total score metric for subscale and composite scores by multiplying values from the item scale metric by the number of items in the scale (i.e., four for subscales, and 12 for composite scores here; see Figure 1).

Results in Table 4 reveal that *G* coefficients across scales range from 0.674 to 0.879 and global *D* coefficients from 0.647 to 0.863. In all instances, *G*, global *D*, and cut-score-specific *D* coefficients for the composite score (see Figure 1) exceed those for subscale scores. This makes sense given that the composite score has three times as many items. For all scales, cut-score-specific *D* coefficients in Figure 1 display the common pattern of being lowest around the scale mean and getting increasingly greater in magnitude as scores deviate from that mean.

In Table 5, we provide observed (lower triangle) and disattenuated (upper triangle) correlations between pairs of Extraversion subscale scores for the $p^{\bullet} \times I^{\circ}$ design. Based on either observed or disattenuated correlations, Assertiveness and Sociability are closely related most strongly (observed $r = .578$, disattenuated $r = .739$), followed by Energy Level and Sociability (observed $r = .526$, disattenuated $r = .718$) and Assertiveness and Energy Level (observed $r = .428$, disattenuated $r = .596$). Of special significance are the noticeable differences in magnitudes between observed (M = 0.511) and disattenuated (M = 0.684) correlations, indicating that constructs measured by these subscales are more strongly related than would otherwise be inferred. This is even most apparent when averaging the coefficients when squared (i.e., M = 0.265 vs. M = 0.472).

Table 5 also includes effective weights for subscale scores based on their variances, covariances, and original weights. Although equal weights were specified initially in the composite formulas for universe score, relative error, and absolute error, the effective weights shown in the table indicate that the proportionate weightings of Assertiveness, Energy Level, and Sociability subscale scores in the composite operationally equal 0.324, 0.265, and 0.410 for universe score, 0.300, 0.344, and 0.356 for relative error, and 0.274, 0.336, and 0.390 for absolute error.

Multivariate GT Analyses for the $p^{\bullet} \times i^{\circ} \times o^{\bullet}$ Design

For $p^{\bullet} \times i^{\circ} \times o^{\bullet}$ multivariate GT designs, seven variance–covariance matrices are estimated: Σ_p, Σ_{pi}, Σ_{po}, $\Sigma_{pio,e}$, Σ_i, Σ_o, and Σ_{io}. Diagonal elements again represent variance components for univariate GT analyses for each individual subscale, and diagonal and off-diagonal elements represent variance and covariance components needed for multivariate GT analyses of composite scores (see Table 2). *G* and global *D* coefficients for subscale and composite scores for the $p^{\bullet} \times I^{\circ} \times O^{\bullet}$ design based on $n_i' = 4$ for each subscale, $n_i' = 12$, for the composite, and $n_o' = 1$ are provided in Table 4; corresponding cut-score-specific *D* coefficients are shown in the bottom graph of Figure 1.

Overall, the same relative magnitudes of *G* and global *D* coefficients observed for the $p^{\bullet} \times I^{\circ}$ design hold for the $p^{\bullet} \times I^{\circ} \times O^{\bullet}$ design, with Extraversion composite scores ($G = .836$, global $D = .821$) having the highest consistency followed, respectively, by Sociability ($G = .771$, global $D = .729$), Assertiveness ($G = .717$, global $D = .709$), and Energy Level ($G = .677$, global $D = .653$). However, the $p^{\bullet} \times I^{\circ} \times O^{\bullet}$ design further allows for partitioning of multiple sources of measurement error, with each source accounting for noteworthy proportions of observed score variance, ranging across scales from 0.069 to 0.171 for specific-factor error, from 0.025 to 0.051 for transient error, and from 0.047 to 0.127 for random-response error (see Table 4). In comparison to subscale scores, composite scores have lower proportions of specific-factor and random-response error, attributable to the composite scale having many more

Table 3
Computation of Extraversion Composite Score Variance Components, Measurement Error, and Score Consistency Indices for $p^\bullet \times i^\circ$ and $p^\bullet \times i^\circ \times o^\bullet$ Designs

$p^\bullet \times i^\circ$ design

$$\hat{\sigma}_c(p) = \sum_{i=1}^{n_s} w_{S_i}^2 \sigma_{S_i}^2(p) + 2\left[\sum_{i=1}^{n_s}\sum_{j>i}^{n_s} w_{S_i} w_{S_j} \sigma_{S_i,S_j}(p)\right]$$
$$= \left(\left(\frac{1}{3}\right)^2 \times .808\right) + \left(\left(\frac{1}{3}\right)^2 \times .572\right) + \left(\left(\frac{1}{3}\right)^2 \times .411\right) + 2 \times \left(\left(\frac{1}{3}\right)^2 \times .503 + \left(\frac{1}{3}\right)^2 \times .414 + \left(\frac{1}{3}\right)^2 \times .289\right) = .467$$

$$\hat{\sigma}_c(pi, e) = \sum_{i=1}^{n_s} w_{S_i}^2 \sigma_{S_i}^2(pi, e)$$
$$= \left(\left(\frac{1}{3}\right)^2 \times .824\right) + \left(\left(\frac{1}{3}\right)^2 \times .694\right) + \left(\left(\frac{1}{3}\right)^2 \times .797\right) = .257$$

$$\hat{\sigma}_c(i) = \sum_{i=1}^{n_s} w_{S_i}^2 \sigma_{S_i}^2(i)$$
$$= \left(\left(\frac{1}{3}\right)^2 \times .218\right) + \left(\left(\frac{1}{3}\right)^2 \times .037\right) + \left(\left(\frac{1}{3}\right)^2 \times .100\right) = .039$$

$$G \text{ coefficient} = \frac{\hat{\sigma}_c^2(p)}{\hat{\sigma}_c^2(p) + \dfrac{\hat{\sigma}_c^2(pi, e)}{n'_{S(i)}}}$$
$$= .467/(.467 + .2574) = .879$$

$$\text{Global } D \text{ coefficient} = \frac{\hat{\sigma}_c^2(p)}{\hat{\sigma}_c^2(p) + \left[\dfrac{\hat{\sigma}_c^2(pi, e)}{n'_{s(i)}} + \dfrac{\hat{\sigma}_{c(i)}^2}{n'_{s(i)}}\right]}$$
$$= .467/(.467 + .2574 + .0394) = .863$$

$p^\bullet \times i^\circ \times o^\bullet$ design

$$\hat{\sigma}_c(p) = \sum_{i=1}^{n_s} w_{S_i}^2 \sigma_{S_i}^2(p) + 2\left[\sum_{i=1}^{n_s}\sum_{j>i}^{n_s} w_{S_i} w_{S_j} \sigma_{S_i,S_j}(p)\right]$$
$$= \left(\left(\frac{1}{3}\right)^2 \times .754\right) + \left(\left(\frac{1}{3}\right)^2 \times .531\right) + \left(\left(\frac{1}{3}\right)^2 \times .421\right) + 2 \times \left(\left(\frac{1}{3}\right)^2 \times .468 + \left(\frac{1}{3}\right)^2 \times .392 + \left(\frac{1}{3}\right)^2 \times .263\right) = .439$$

$$\hat{\sigma}_c(pi) = \sum_{i=1}^{n_s} w_{S_i}^2 \sigma_{S_i}^2(pi)$$
$$= \left(\left(\frac{1}{3}\right)^2 \times .505\right) + \left(\left(\frac{1}{3}\right)^2 \times .366\right) + \left(\left(\frac{1}{3}\right)^2 \times .427\right) = .144$$

$$\hat{\sigma}_c(po) = \sum_{i=1}^{n_s} w_{S_i}^2 \sigma_{S_i}^2(po) + 2\left[\sum_{i=1}^{n_s}\sum_{j>i}^{n_s} w_{S_i} w_{S_j} \sigma_{S_i,S_j}(po)\right]$$
$$= \left(\left(\frac{1}{3}\right)^2 \times .036\right) + \left(\left(\frac{1}{3}\right)^2 \times .038\right) + \left(\left(\frac{1}{3}\right)^2 \times .015\right) + 2\left[\left(\left(\frac{1}{3}\right)^2 \times .023\right) + \left(\left(\frac{1}{3}\right)^2 \times .026\right) + \left(\left(\frac{1}{3}\right)^2 \times .020\right)\right] = .025$$

$$\hat{\sigma}_c(pio, e) = \sum_{i=1}^{n_s} w_{S_i}^2 \sigma_{S_i}^2(pio, e) = \left(\left(\frac{1}{3}\right)^2 \times .249\right) + \left(\left(\frac{1}{3}\right)^2 \times .323\right) + \left(\left(\frac{1}{3}\right)^2 \times .315\right) = .098$$

$$\hat{\sigma}_c(i) = \sum_{i=1}^{n_s} w_{S_i}^2 \sigma_{S_i}^2(i)$$
$$= \left(\left(\frac{1}{3}\right)^2 \times .221\right) + \left(\left(\frac{1}{3}\right)^2 \times .036\right) + \left(\left(\frac{1}{3}\right)^2 \times .094\right) = .039$$

$$\hat{\sigma}_c(o) = \sum_{i=1}^{n_s} w_{S_i}^2 \sigma_{S_i}^2(o) + 2\left[\sum_{i=1}^{n_s}\sum_{j>i}^{n_s} w_{S_i} w_{S_j} \sigma_{S_i,S_j}(o)\right]$$
$$= \left(\left(\frac{1}{3}\right)^2 \times .000\right) + \left(\left(\frac{1}{3}\right)^2 \times .000\right) + \left(\left(\frac{1}{3}\right)^2 \times .000\right) + 2\left[\left(\left(\frac{1}{3}\right)^2 \times .000\right) + \left(\left(\frac{1}{3}\right)^2 \times .000\right) + \left(\left(\frac{1}{3}\right)^2 \times .000\right)\right] = .000$$

$$\hat{\sigma}_c(io) = \sum_{i=1}^{n_s} w_{S_i}^2 \sigma_{S_i}^2(io)$$
$$= \left(\left(\frac{1}{3}\right)^2 \times -.001\right) + \left(\left(\frac{1}{3}\right)^2 \times -.001\right) + \left(\left(\frac{1}{3}\right)^2 \times .001\right) = .000$$

$$\text{Specific-factor error} = \frac{\frac{\hat{\sigma}_c^2(pi)}{n'_{s(i)}}}{\hat{\sigma}_c^2(p) + \left[\frac{\hat{\sigma}_c^2(pi)}{n'_{s(i)}} + \frac{\hat{\sigma}^2(po)}{n'_{s(o)}} + \frac{\hat{\sigma}_{c(pio,e)}^2}{n'_{s(i)}n'_{s(o)}}\right]}$$
$$= (.1444)/(.439 + .1444 + .0251 + .0984) = .069$$

$$\text{Transient error} = \frac{\frac{\hat{\sigma}^2(po)}{n'_{s(o)}}}{\hat{\sigma}_c^2(p) + \left[\frac{\hat{\sigma}_c^2(pi)}{n'_{s(i)}} + \frac{\hat{\sigma}^2(po)}{n'_{s(o)}} + \frac{\hat{\sigma}_{c(pio,e)}^2}{n'_{s(i)}n'_{s(o)}}\right]}$$
$$= (.0251)/(.439 + .1444 + .0251 + .0984) = .048$$

$$\text{Random-response error} = \frac{\frac{\hat{\sigma}_{c(pio,e)}^2}{n'_{s(i)}n'_{s(o)}}}{\hat{\sigma}_c^2(p) + \left[\frac{\hat{\sigma}_c^2(pi)}{n'_{s(i)}} + \frac{\hat{\sigma}^2(po)}{n'_{s(o)}} + \frac{\hat{\sigma}_{c(pio,e)}^2}{n'_{s(i)}n'_{s(o)}}\right]}$$
$$= (.0984)/(.439 + .1444 + .0251 + .0984) = .047$$

$$G \text{ coefficient} = \frac{\hat{\sigma}_c^2(p)}{\hat{\sigma}_c^2(p) + \left[\frac{\hat{\sigma}_c^2(pi)}{n'_{s(i)}} + \frac{\hat{\sigma}^2(po)}{n'_{s(o)}} + \frac{\hat{\sigma}^2_{c(pio,e)}}{n'_{s(i)}n'_{s(o)}}\right]}$$

$$= .439/(.439 + .1444 + .0251 + .0984) = .836$$

$$\text{Global } D \text{ coefficient} = \frac{\hat{\sigma}_c^2(p)}{\hat{\sigma}_c^2(p) + \left[\frac{\hat{\sigma}_c^2(pi)}{n'_{s(i)}} + \frac{\hat{\sigma}^2(po)}{n'_{s(o)}} + \frac{\hat{\sigma}^2_{c(pio,e)}}{n'_{s(i)}n'_{s(o)}} + \frac{\hat{\sigma}_c^2(i)}{n'_{s(i)}} + \frac{\hat{\sigma}^2(o)}{n'_{s(o)}} + \frac{\hat{\sigma}^2_{c(io)}}{n'_{s(i)}n'_{s(o)}}\right]}$$

$$= .439/(.439 + .1444 + .0251 + .0984 + .0394 + .0001 + .0004) = .821$$

Note. Results obtained directly from formulas in this table may differ slightly from the answers given due to rounding. The final values provided come directly from output produced by *mGENOVA* (Brennan, 2001b).

items. Across the three sources of measurement error, transient ($M = 0.038$) is lowest on average followed, respectively, by random-response ($M = 0.112$) and specific-factor ($M = 0.145$). This pattern implies that score consistency would likely be better improved by increasing items than by increasing occasions. Because the $p^\bullet \times I^\circ \times O^\bullet$ design includes additional sources of measurement error, G, global D, and cut-score-specific D coefficients are typically lower than corresponding indices in the $p^\bullet \times I^\circ$ design (see Table 4 and Figure 1).

Patterns of correlation coefficients for the $p^\bullet \times I^\circ \times O^\bullet$ design in Table 5 parallel those for the $p^\bullet \times I^\circ$ design. Assertiveness and Sociability scores are most highly correlated (observed $r = .550$, disattenuated $r = .739$), followed by Energy Level and Sociability (observed $r = .503$, disattenuated $r = .696$) and Assertiveness and Energy Level (observed $r = .388$, disattenuated $r = .557$). Differences between observed ($M_r = 0.480$, $M_{r^2} = 0.235$) and disattenuated ($M_r = 0.664$, $M_{r^2} = 0.447$) correlations again reveal significant underestimation of relationships between constructs when using observed score correlations. The ordering of effective weights for subscales in the current design shown in the Table 5 coincides with that from the previous design except for Assertiveness's effective weight for relative error now being greater than that for Energy Level.

Further Applications of Multivariate GT Designs

Changing Measurement Procedures and Universes of Generalization

Two important applications within decision studies mentioned earlier are to estimate how GT indices of score consistency and measurement error are affected when changing a measurement procedure and when restricting universes of generalization to a lesser number of facets. For the $p^\bullet \times I^\circ \times O^\bullet$ multivariate GT design, we could determine how indices of score consistency and measurement error change when altering numbers of items and/or occasions, changing weights for subscale scores, and/or restricting the universe of generalization to just items or just occasions.

In Table 6, we provide estimates of composite score G coefficients, and proportions of specific-factor, transient, and random-response error, for (a) numbers of items within each subscale equaling 4, 6, and 8; (b) number of occasions equaling 1, 2, and 3; and (c) initial weights for Assertiveness, Energy Level, and Sociability subscale scores equaling .33/.33/.33, .25/.25/.5, and .4/.4/.2. The subscale weights just described are intended to contrast situations in which: (a) the three subscales are initially weighted equally, (b) Sociability is weighted twice as heavily as Assertiveness and Energy Level, and (c) Assertivenes and Energy Level are weighted twice as heavily as Sociability.

As would be expected, adding items to subscales increases G and global D coefficients by reducing specific-factor and random-response error, with score consistency increasing more when going from four to six items per subscale than when going from six to eight (see Table 6). Adding occasions also increases G and global D coefficients, but in this case, by reducing transient and random-response error, with those coefficients increasing more when going from one to two than from two to three occasions. This trend of diminishing gains in consistency would continue with similar incremental increases to numbers of items or occasions. G and global D coefficients do not vary as noticeably when changing weights, but nevertheless show that weighting the most reliable subscale (Sociability) equally or more heavily than the other subscales yields higher composite score consistency than when giving greater weight to less reliable subscales.[5]

Restricted universes for the $p^\bullet \times I^\circ \times O^\bullet$ design would include generalizing just across items or just across occasions. As noted earlier, when such restrictions occur, the excluded source of measurement error is treated as universe score variance. Transient error is confounded with person variance when generalizing just across items, and specific-factor error is confounded with person variance when generalizing just across occasions. As a result, reliability indices will generally be greater in the restricted universe. Partitioning of measurement error variance in Table 4 verifies these relationships

[5] Vispoel, Lee, & Hong (in press) reported a similar overall pattern of relationships for the BFI-2 Conscientiousness composite scale and its three nested subscales (Organization, Productiveness, and Responsibility). However, an important point to stress here is that weights intended to maximize reliability do not necessarily maximize validity. For example, in a simulation study, Baldwin (2015) demonstrated that composite scores based on expert judgments of weights can exceed the validity of composite scores using weights that maximize reliability when validity is defined as the correlation between observed and true composite scores.

Table 4
Partitioning of Variance for $p^\bullet \times I^\circ$ and $p^\bullet \times I^\circ \times O^\bullet$ Designs

Design/scale	Index					
	G coef/US	SFE	TE	RRE	Total error	Global *D* coef
$p^\bullet \times I^\circ$ design						
Extraversion	.879 (.872, .886)				.121 (.114, .128)	.863 (.854, .870)
Assertiveness	.767 (.738, .795)				.233 (.205, .262)	.757 (.727, .785)
Energy level	.674 (.634, .710)				.326 (.290, .366)	.647 (.606, .684)
Sociability	.797 (.777, .817)				.203 (.183, .223)	.756 (.733, .777)
Mean subscales	.746				.254	.720
Mean all scales	.779				.221	.756
$p^\bullet \times I^\circ \times O^\bullet$ design						
Extraversion	.836 (.822, .851)	.069 (.065, .072)	.048 (.032, .064)	.047 (.042, .052)	.164 (.149, .178)	.821 (.806, .834)
Assertiveness	.717 (.686, .748)	.123 (.110, .137)	.051 (.013, .088)	.109 (.091, .127)	.283 (.252, .314)	.708 (.676, .738)
Energy level	.677 (.641, .714)	.172 (.155, .188)	.025 (−.021, .069)	.127 (.106, .148)	.323 (.286, .359)	.652 (.616, .686)
Sociability	.771 (.747, .795)	.129 (.119, .140)	.036 (.008, .065)	.064 (.050, .077)	.229 (.205, .253)	.729 (.705, .751)
Mean subscales	.722	.141	.037	.100	.278	.696
Mean all scales	.750	.123	.040	.086	.250	.727

Note. $n'_i = 4$ per subscale in both designs above, and $n'_o = 1$ for the $p^\bullet \times I^\circ \times O^\bullet$ design. All analyses are based on observed scores. Values within parentheses for score consistency and measurement error indices in Columns 2–7 represent 80% Monte Carlo-based confidence interval limits obtained from the *semTools* package in R (Jorgensen et al., 2022). *G* coefficients are equivalent to proportions of universe score variance. *G* coef = generalizability coefficient; US = universe score; SFE = specific-factor error; TE = transient error; RRE = random-response error; *D* coef = dependability coefficient.

within all models represented. *G* and global *D* coefficients are uniformly lower when generalizing across both items and occasions than when generalizing just across one or the other. For example, within the $p^\bullet \times I^\circ \times O^\bullet$ design with equal initial weights, $n'_i = 4$ per subscale, and $n'_o = 1$, the *G* coefficient for Extraversion composite scores equals 0.836 when generalizing across both items and occasions, compared to 0.884 (=0.836 + 0.048) when generalizing just across items and 0.905 (=0.836 + 0.069) when generalizing just across occasions (see Equations 34 and 36).

Estimating Score Consistency for Partially Crossed Designs

Examples considered thus far are based on fully crossed univariate designs for each subscale in which respondents completed all items across all occasions. However, there are situations in which different respondents might answer different items or different raters might evaluate performances of different individuals. If we assume that items (or raters) are randomly assigned to persons, we can use results from the present designs to estimate score consistency for such partially crossed designs. For the $p^\bullet \times I^\circ$ and $p^\bullet \times I^\circ \times O^\bullet$ designs illustrated here, we will assume that two items within each subscale are randomly administered to each respondent. Because we initially assumed that the original four items from each subscale were randomly selected from larger domains of interest, we can use the same variance and covariance components from the original *G* study and formulas presented earlier to estimate score consistency and measurement error for partially crossed designs by specifying that number of items equals two. We provide these results in Table 7. Note that indices for score consistency are uniformly lower, and those for measurement error are uniformly higher than those reported in Table 4 for the same designs with four items.

Evaluating Subscale Score Viability

In Table 8, we provide PRMSE indices for Extraversion subscale and composite scores and their corresponding VARs. These indices are based on *G* coefficients for (a) the $p^\bullet \times I^\circ$ design with $n'_i = 4$ per subscale, (b) $p^\bullet \times O^\bullet$ designs with $n'_o = 1$ or ~~$n'_o =$~~ 2, and (c) $p^\bullet \times I^\circ \times O^\bullet$ designs with $n'_i = 4$ per subscale and $n'_o = 1$ or 2. *G* coefficients for the single occasion $p^\bullet \times I^\circ$ design are equivalent to conventional alpha coefficients, and *G* coefficients for $p^\bullet \times O^\bullet$ design with $n'_o = 1$ are analogous to conventional test–retest coefficients.[6] In all instances, except for the Sociability subscale within the $p^\bullet \times I^\circ \times O^\bullet$ design with $n'_o = 1$, subscale viability is supported. VARs range from ~~.970~~ to 1.498 and reveal that the degree of subscale added value varies with the GT design, magnitude of subscale intercorrelations, and corresponding *G* coefficients. Added value is lower for the Sociability subscale than for the Assertiveness and Energy Level subscales and strongest across subscales when estimating reliability using *G* coefficients from the $p^\bullet \times O^\bullet$ design in which results are pooled across two occasions (i.e., $n'_o = 2$).

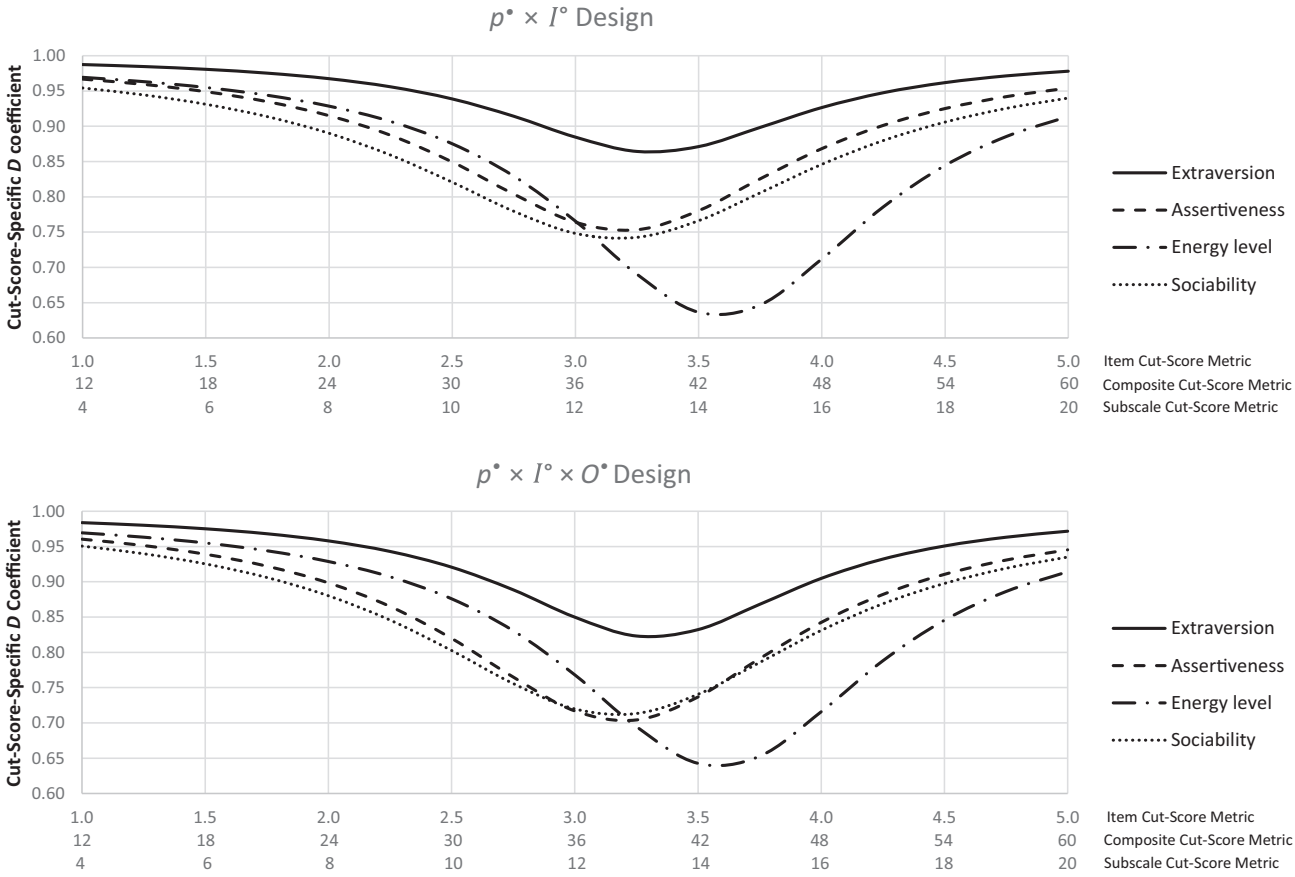
Software for Analyzing Multivariate GT Designs

All multivariate GT designs illustrated here can be analyzed using the *mGENOVA* package. This package and accompanying manual are available at no cost from the following website: http://www.education.uiowa.edu/centers/casma/computer-programs. Although *mGENOVA* has historically served as the gold standard for performing multivariate GT analyses, the *gtheory* (Moore, 2016), *glmmTMB* (Brooks et al., 2017), and *lavaan* (Rosseel, 2012) packages in R also

[6] G coefficients for univariate $p \times o$ designs, based on two observed occasions and using $n'_o = 1$ in relevant equations, are equivalent to conventional test-retest coefficients when occasion score variances are equal but are lower in other instances. This occurs because the denominator for test-retest coefficients includes a *geometric mean* that is less than or equal to the *arithmetic mean* within the corresponding denominator for G coefficients (see Vispoel et al., 2018a, pp. 7–8). In general, differences between test–retest and associated G coefficients tend to be small unless occasion score variances noticeably differ. To five decimal places for the data reported here, test–retest coefficients for Assertiveness, Energy Level, and Sociability equal 0.84027, 0.84886, and 0.90063 versus corresponding G coefficients that equal 0.84026, 0.84872, and 0.90000.

Figure 1

Multivariate GT Designs Cut-score-Specific D Coefficients for Extraversion Composite and Subscale Scores



Note. GT= generalizability theory. Minimum cut-score-specific *D* coefficient values in the figure may vary slightly from corresponding global *D* coefficient values due to corrections for bias (see online supplemental materials).

can be used for such purposes with many designs.[7] The *gtheory* package by itself can only perform multivariate GT analyses for facets nested within subscales. This would work for $p^\bullet \times i^\circ$ and $p^\bullet \times i^\circ \times o^\circ$ designs but not for $p^\bullet \times i^\bullet$, $p^\bullet \times i^\circ \times o^\bullet$, and $p^\bullet \times i^\bullet \times o^\bullet$ designs. Jiang et al. (2020) showed how selected multivariate GT designs can be analyzed using the *glmmTMB* package and discussed advantages and disadvantages of this approach compared to *mGENOVA*. Vispoel, Lee, and Hong (in press) further demonstrated that multivariate GT designs also can be analyzed using structural equation modeling packages such as *lavaan* in R. When using unweighted least squares estimation with *lavaan*, they obtained essentially the same results as those from the *mGENOVA* package for a variety of multivariate designs.

The *glmmTMB* and *lavaan* packages in R will produce estimates of variance and covariance components for the multivariate GT analyses discussed here that can be inserted into relevant formulas to derive *G* coefficients, *D* coefficients, effective weights, proportions of universe score and measurement error variance, and other indices. Additional code can be added within R to compute those indices directly and to implement procedures for handling missing data and constructing Monte Carlo-based confidence intervals that are unavailable within *mGENOVA* (see our online supplemental materials).

In Figures 2 and 3, we depict examples of structural equation models for analyzing multivariate $p^\bullet \times i^\circ$ and $p^\bullet \times i^\circ \times o^\bullet$ designs for composite and subscale scores within the BFI-2 Extraversion domain. The multivariate $p^\bullet \times i^\bullet$ model would be the same as the $p^\bullet \times i^\circ$ model in Figure 2 except that item factors with common stems and corresponding uniquenesses would be allowed to covary/correlate across subscales. The multivariate $p^\bullet \times i^\circ \times o^\circ$ and $p^\bullet \times i^\bullet \times o^\bullet$ models would differ from the $p^\bullet \times i^\circ \times o^\bullet$ model shown in Figure 3 in that occasion/transient error factors would be uncorrelated across subscales in the $p^\bullet \times i^\circ \times o^\circ$ design, and item factors with common stems and corresponding uniquenesses would be allowed to covary/correlate across subscales in the $p^\bullet \times i^\bullet \times o^\bullet$ design (see Vispoel, Lee, & Hong, in press and our online supplemental materials for further details).

[7] *mGENOVA* is restricted to one and two measurement error facets in multivariate designs, whereas code for analyzable designs within the R packages mentioned here could be adapted to handle additional facets. In their online supplemental materials, Vispoel et al. (2019) provide partial code within M*plus* (Muthén & Muthén, 2017) and the *lavaan* package in R (Rosseel, 2012) for analyzing univariate and multivariate GT designs with three measurement error facets (i.e., tasks, raters, and occasions).

Table 5
Correlation Coefficients and Effective Weights for $p^\bullet \times I^\circ$ and $p^\bullet \times I^\circ \times O^\bullet$ Designs

	Subscale in $p^\bullet \times I^\circ$ design			Subscale in $p^\bullet \times I^\circ \times O^\bullet$ design		
Subscale	Assertiveness	Energy level	Sociability	Assertiveness	Energy level	Sociability
Correlation coefficients						
Assertiveness	—	0.596	0.739	—	0.557	0.739
Energy level	0.428	—	0.718	0.388	—	0.696
Sociability	0.578	0.526	—	0.550	0.503	—
Effective weights						
Universe score	0.324	0.265	0.410	0.319	0.272	0.408
Relative error	0.300	0.344	0.356	0.327	0.320	0.353
Absolute error	0.274	0.336	0.390	0.304	0.314	0.382

Note. Observed score correlation coefficients are in the lower triangle of the matrices, and disattenuated correlations are in the upper triangle. $n_i' = 4$ per subscale in both designs above, and $n_o' = 1$ for the $p^\bullet \times I^\circ \times O^\bullet$ design. All analyses are based on observed scores.

The models shown in Figures 2 and 3 visually illustrate the point made at the outset of this tutorial that multivariate GT models consist of random-effects univariate models for each subscale linked together via covariance components. Factor loadings for both models are set equal to 1, and variance and covariance components involving relative error are directly estimated. Jorgensen (2021) (also see Vispoel, Hong, et al., in press; Vispoel, Hong, & Lee, 2023; Vispoel, Lee, Chen, & Hong, 2023a) demonstrated that variance components ~~involving absolute error~~ can be obtained indirectly by imposing effect coding and other constraints within each subscale. When using effect coding (see Little et al., 2006), factor loadings are constrained to average 1, and the sum of item intercepts to equal zero. Under these constraints within the $p \times i$ designs, the σ_i^2 component for a given subscale can be estimated using the following equation:

$$\hat{\sigma}_i^2 = \frac{1}{n_i - 1}\sum_{1}^{n_i}(\text{Intercept}_i)^2. \tag{40}$$

Within the $p \times i \times o$ designs, effect coding constraints are again imposed and sums for item and occasion factor means are set equal to zero. Once these constraints are imposed, the remaining variance components for those designs can be obtained from the following equations:

$$\hat{\sigma}_i^2 = \frac{1}{n_i - 1}\sum_{1}^{n_i}(\text{Item factor mean}_i)^2, \tag{41}$$

$$\hat{\sigma}_o^2 = \frac{1}{n_o - 1}\sum_{1}^{n_o}(\text{Occasion factor mean}_o)^2, \tag{42}$$

$$\hat{\sigma}_{io}^2 = \frac{1}{(n_i \times n_o) - 1}\sum_{1}^{n_i \times n_o}(\text{Intercept}_{io})^2. \tag{43}$$

Comparing Results Across Packages

In our online supplemental materials, we provide code and output from the *gtheory* package for the $p^\bullet \times i^\circ$ design and from the *mGENOVA*, *glmmTMB*, and *lavaan* packages for the present $p^\bullet \times i^\circ$ and $p^\bullet \times i^\circ \times o^\bullet$ designs. Table 9 includes score consistency and variance partitioning indices for these designs using those programs. Note that results from *mGENOVA*, *gtheory*, and *lavaan* vary by no more than 0.001, but differ to a greater extent with those yielded by *glmmTMB*. Jiang et al. (2020) attributed such discrepancies to functional differences in algorithms used to estimate variance and covariance components. They noted, for example,

Table 6
Partitioning of Variance for Altered $p^\bullet \times I^\circ \times O^\bullet$ Designs

			Index					
Initial weights	# of items	# of occasions	*G* coef/US	SFE	TE	RRE	Total error	Global *D* coef
.33/.33/.33	4	1	0.836	0.069	0.048	0.047	0.164	0.821
.33/.33/.33	6	1	0.870	0.048	0.050	0.032	0.130	0.859
.33/.33/.33	8	1	0.887	0.036	0.051	0.025	0.112	0.879
.33/.33/.33	4	2	0.878	0.072	0.025	0.025	0.122	0.861
.33/.33/.33	4	3	0.893	0.073	0.017	0.017	0.107	0.876
.25/.25/.5	4	1	0.837	0.074	0.046	0.043	0.163	0.815
.4/.4/.2	4	1	0.820	0.074	0.049	0.057	0.180	0.807

Note. *G* coefficients are equivalent to proportions of universe score variance. All analyses are based on observed scores. *G* coef = generalizability coefficient; US = universe score; SFE = specific-factor error; TE = transient error; RRE = random-response error; *D* coef = dependability coefficient.

Table 7
Partitioning of Variance for Partially Crossed $p^\bullet \times I^\circ$ and $p^\bullet \times I^\circ \times O^\bullet$ Designs With $n'_i = 2$ per Subscale, and $n'_o = 1$

Design/scale	Index					
	G coef/US	SFE	TE	RRE	Total error	Global *D* coef
$p^\bullet \times I^\circ$ design						
Extraversion	.784 (.773, .795)				.216 (.205, .227)	.759 (.746, .770)
Assertiveness	.622 (.585, .659)				.378 (.341, .415)	.609 (.571, .645)
Energy level	.508 (.465, .551)				.492 (.449, .535)	.478 (.435, .519)
Sociability	.662 (.635, .690)				.338 (.310, .365)	.608 (.578, .636)
Mean subscales	.598				.402	.565
Mean all scales	.644				.356	.613
$p^\bullet \times I^\circ \times O^\bullet$ design						
Extraversion	.750 (.737, .762)	.123 (.116, .130)	.043 (.029, .058)	.084 (.076, .092)	.250 (.238, .263)	.725 (.712, .736)
Assertiveness	.582 (.558, .606)	.200 (.178, .222)	.041 (.011, .072)	.177 (.149, .204)	.418 (.394, .442)	.570 (.545, .593)
Energy level	.522 (.495, .548)	.264 (.239, .290)	.019 (−.016, .053)	.195 (.164, .226)	.478 (.452, .505)	.492 (.466, .517)
Sociability	.646 (.627, .666)	.217 (.199, .234)	.031 (.006, .054)	.107 (.085, .128)	.354 (.334, .373)	.590 (.571, .607)
Mean subscales	.583	.227	.030	.159	.417	.551
Mean all scales	.625	.201	.034	.141	.375	.594

Note. All analyses are based on observed scores. Values within parentheses for score consistency and measurement error indices in columns 2–7 represent 80% Monte Carlo-based confidence interval limits obtained from the *semTools* package in R (Jorgensen et al., 2022). *G* coefficients are equivalent to proportions of universe score variance. *G* coef = generalizability coefficient; US = universe score; SFE = specific-factor error; TE = transient error; RRE = random-response error; *D* coef = dependability coefficient.

that the *glmmTMB* package uses automatic differentiation to estimate model gradients and Laplace approximations to handle random effects, whereas *mGENOVA* uses an ANOVA-based algorithm targeting expected mean squares.

Handling Missing Data and Scale Coarseness

General procedures for handling missing responses can be applied to data before using any of these packages. Traditional procedures for self-report measures include eliminating cases with any missing data (i.e., listwise deletion) or setting a cutoff for number of completed items within a subscale (e.g., 75%) and replacing the score for a missing item with the person's mean or median for completed items or with the grand mean or median for the item across persons. However, such procedures can produce biased results and have been largely supplanted by multiple imputation and other techniques (Engers, 2022). Examples of packages within R that replace missing with imputed values include *Amelia II* (Honaker et al., 2011, 2022), *Hmisc* (Harrell & Dupont, 2022), *mi*. (Su et al., 2022), *mice* (van

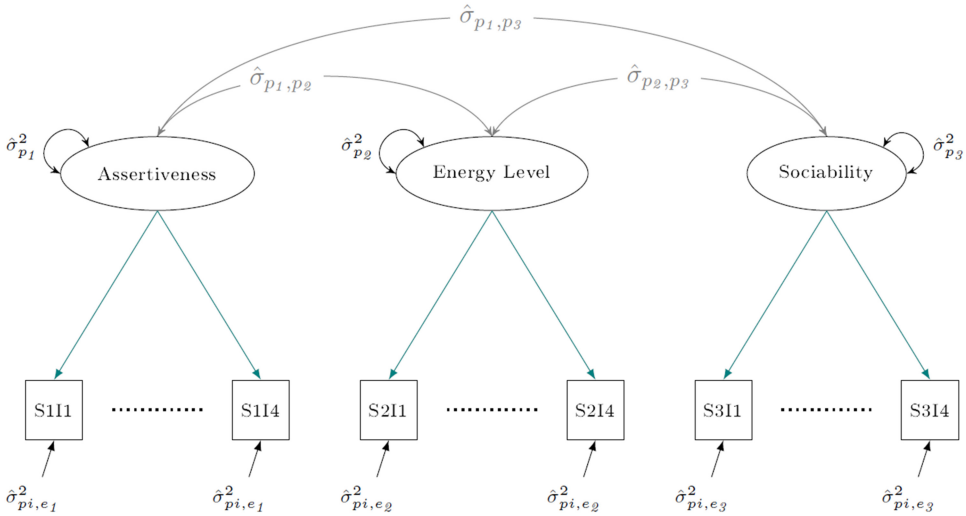
Table 8
Scale Viability Indices for Selected Multivariate GT Designs

Design/subscale	Index for ULS estimation			Index for WLSMV estimation		
	PRMSE(S)	PRMSE(C)	VAR	PRMSE(S)	PRMSE(C)	VAR
$p^\bullet \times I^\circ \times O^\bullet$ design ($n'_o = 1$)						
Assertiveness	.717	.701	1.023	.778	.690	1.128
Energy level	.677	.651	1.040	.749	.635	1.180
Sociability	.771	.795	0.970	.833	.790	1.054
$p^\bullet \times I^\circ \times O^\bullet$ design ($n'_o = 2$)						
Assertiveness	.835	.682	1.224	.875	.678	1.291
Energy level	.807	.625	1.291	.857	.619	1.384
Sociability	.871	.788	1.105	.909	.787	1.155
$p^\bullet \times O^\bullet$ ($n'_o = 1$)						
Assertiveness	.840	.682	1.233	.843	.681	1.238
Energy level	.849	.620	1.369	.861	.618	1.393
Sociability	.900	.787	1.144	.906	.787	1.151
$p^\bullet \times O^\bullet$ ($n'_o = 2$)						
Assertiveness	.913	.675	1.353	.915	.675	1.356
Energy level	.918	.613	1.498	.925	.612	1.511
Sociability	.947	.786	1.205	.951	.786	1.210
$p^\bullet \times I^\circ$ design						
Assertiveness	.767	.680	1.129	.833	.672	1.240
Energy level	.674	.631	1.067	.809	.607	1.333
Sociability	.797	.770	1.035	.876	.767	1.142

Note. $n'_i = 4$ per subscale in all designs. GT = generalizability theory; ULS = unweighted least squares estimation; WLSMV = weighted least squares with adjusted means and variances (i.e., diagonally weighted least squares in R); PRMSE(S) = proportional reduction in mean-squared error for the subscale scores; PRMSE(C) = proportional reduction in mean-squared error for composite scores; VAR = value-added ratio = PRMSE(S)/PRMSE(C).

Figure 2

Structural Equation Model for the Single-Occasion $p^{\bullet} \times i^{\circ}$ Multivariate Design Using BFI-2 Extraversion Subscale Scores



Note. In the structural equation model above, factor loadings are set equal to 1, persons scores covary across subscales, and uniquenesses ($\hat{\sigma}^2_{pi,e}$) are set equal within but not across subscales. BFI-2 = Big Five Inventory-2; S = subscale; I = item. See the online article for the color version of this figure.

Buuren & Groothuis-Oudshoorn, 2022), *mitools* (Lumley, 2022), *missForest* (Stekhoven, 2022), and *mitml* (Grund et al., 2023). Although there were no missing values within the present data set, we provide code for using the *mice* (multiple imputation by chained equations) package in R to impute predicted values for missing responses in our online supplemental materials.

As an alternative when using structural equation models to perform multivariate GT analyses under the assumption that omitted responses are missing at random, the preanalysis multiple imputation procedures described above can be avoided altogether by using full information maximum likelihood or Bayesian estimation procedures (see, e.g., Engers, 2022; Garnier-Villarreal & Jorgensen, 2020; Jiang & Skorupski, 2018; Merkle, 2011; Merkle & Rosseel, 2018). Accordingly, we include examples of applying these approaches in our online supplemental materials. Jiang et al. (2018) also demonstrated via simulated data that bias in estimating *G* coefficients within two-facet univariate GT designs attributable to missing data could be reduced by further altering structural equation model analyses to include relevant auxiliary information. We refer readers to their article for details regarding such procedures and note that they can be readily extended to multivariate GT analyses when such information is available. If R is used to conduct structural equation modeling analyses, the *anxiliary*, *BootMiss-class*, *bsMissBoot*, and other procedures within the *semTools* package can be coded directly and linked to *lavaan* to handle missing data using auxiliary information, multiple imputation, bootstrapping, and related procedures (see Jorgensen et al., 2022 for further details).

In addition to handling missing data directly, GT-based structural equation models allow for conducting univariate and multivariate GT analyses on continuous latent response variable metrics that correct for scale coarseness effects resulting from limited numbers of response options and/or unequal intervals between those options using robust diagonally weighted least squares, paired maximum likelihood, or other appropriate estimators (see, e.g., Ark, 2015; Jorgensen, 2021; Vispoel et al., 2019; Vispoel, Lee, & Hong, in press; Vispoel, Hong, & Lee, 2023; Vispoel, Lee, Chen, & Hong, 2023a).[8] Together, such corrections for scale coarseness and missing data within structural equation modeling packages represent clear advantages over *mGENOVA*, which permits neither.

In Table 10, we provide continuous latent response variable results for the present $p^{\bullet} \times i^{\circ}$ and $p^{\bullet} \times i^{\circ} \times o^{\bullet}$ designs using robust diagonally weighted least squares estimation (WLSMV within *lavaan*) to produce the same indices reported in Table 4 for observed scores using *mGENOVA*. Note that correcting for scale coarseness yields uniformly higher *G* and global *D* coefficients in both designs, lower proportions of overall error in the $p^{\bullet} \times i^{\circ}$ design, and lower proportions of specific-factor, random-response, and overall error in the $p^{\bullet} \times i^{\circ} \times o^{\bullet}$ design. *G* and global *D* coefficients yielded by WLSMV estimates in *lavaan* can be interpreted as upper bounds for score consistency that might be achieved by increasing responses options within the corresponding observed score designs (Jorgensen, 2021; Vispoel et al., 2019).

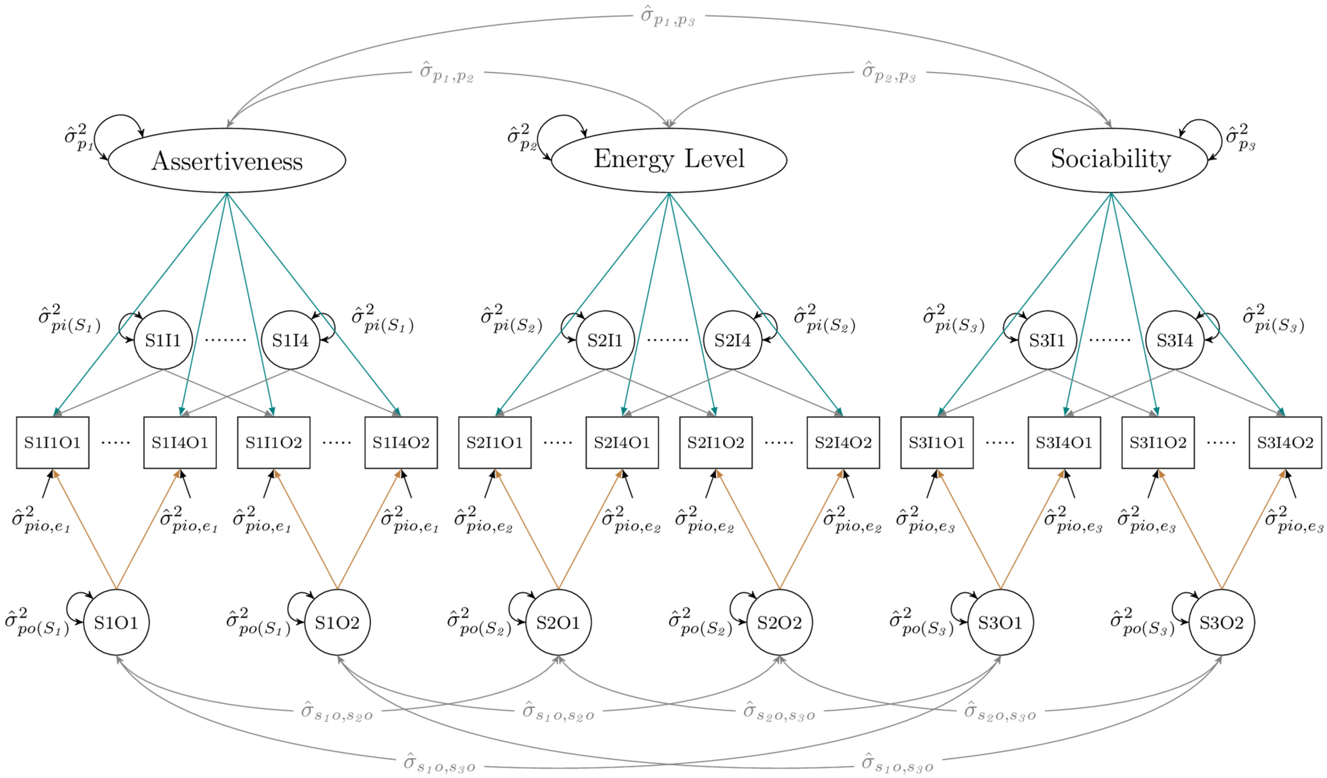
Confidence Intervals

Cronbach et al. (1972), Brennan (2001a), and others have long underscored the importance of gauging sampling variability in GT parameter estimates but procedures to do so are unavailable within most packages for doing GT analyses (including *mGENOVA*) or are limited to methods governed by overly

[8] Diagonally weighted least squares and other forms of estimation also can be used within generalized linear models to correct for scale coarseness (see, e.g., Agresti, 2010).

Figure 3

Structural Equation Model for the Two-Occasion $p^\bullet \times i^\circ \times o^\bullet$ Multivariate Design Using BFI-2 Extraversion Subscale Scores



Note. In the structural equation model above, all factor loadings are set equal to 1; person scores covary across subscales; item factor variances ($\hat{\sigma}^2_{pi}$) and uniquenesses ($\hat{\sigma}^2_{pio,e}$) are set equal within but not across subscales; occasion factor variances ($\hat{\sigma}^2_{po}$) are set equal across occasions but not equal across subscales; and covariances among occasion factors can differ across subscales within an occasion but are set equal across occasions. BFI-2 = Big Five Inventory-2; S = subscale; I = item; O = occasion. See the online article for the color version of this figure.

restrictive assumptions. Creation of the recent *semTools* package in R (Jorgensen et al., 2022) solves this problem by producing Monte Carlo-based confidence intervals for a wide variety of indices computed within or outside of R if an asymptotic sampling covariance matrix of variance-component parameters is available (see our online supplemental materials). Preacher and Selig (2012) noted that the Monte Carlo method for deriving confidence intervals produced results comparable to bootstrapping and distribution of product methods while being less computationally intensive and requiring only summary data.

Confidence intervals for GT parameters provide a means for evaluating the effects for persons, contributions to measurement error, differences in absolute levels of scores, and trustworthiness of score consistency indices by noting whether zero or other targeted parameter values fall within the limits of the interval. In accordance with examples in Brennan (2001a), we provide 80% confidence intervals in Tables 4, 7, and 10 for key GT parameters, including *G* coefficients, global *D* coefficients, and proportions of measurement error variance for BFI-2 Extraversion composite and subscale scores derived from the *semTools* package in R on both observed score and continuous latent response variable metrics. Across these analyses, confidence intervals fail to capture zero for most proportions of measurement error, thereby highlighting the importance of taking those sources of measurement error (overall, specific-factor, transient, and random-response) into account. The only exceptions are with proportions of transient error for the Energy Level subscale within the observed score analyses summarized in Tables 4 and 7. Although well above zero in all designs shown in Tables 4, 7, and 10, lower limits of confidence intervals for *G* and global *D* coefficients are greater for composite than for subscale scores and greater on continuous latent response variable than on observed score metrics.

Alternative Ways to Assess Score Consistency for Composite Scores in GT Designs

Although we emphasize derivation of score consistency indices for composite scores using multivariate GT here, such indices also can be obtained using alternative GT methods. These alternatives include (a) analyzing item scores within a bifactor GT framework in which universe score variance is partitioned into two sources: general (common explained variance across all items) and group (unrelated explained variance unique to each subscale; see, e.g., Vispoel & Lee, 2023; Vispoel, Lee, Chen, & Hong, 2023b, 2023c; Vispoel et al., 2022, 2023), (b) replacing items with parallel splits balanced for subscale representation, item phrasing, and overall statistical characteristics of

Table 9
Partitioning of Observed Score Variance From Alternative Packages for $p^\bullet \times i^\circ$ and $p^\bullet \times i^\circ \times o^\bullet$ Designs

	Index					
Design/program/scale	*G* coef/US	SFE	TE	RRE	Total error	Global *D* coef
$p^\bullet \times I^\circ$ design						
mGENOVA						
Extraversion	.879				.121	.863
Assertiveness	.767				.233	.758
Energy level	.674				.326	.647
Sociability	.797				.203	.756
gtheory package in R						
Extraversion	.879				.121	.863
Assertiveness	.767				.233	.758
Energy level	.674				.326	.647
Sociability	.797				.203	.756
glmmTBM package in R						
Extraversion	.880				.120	.867
Assertiveness	.767				.233	.760
Energy level	.674				.326	.654
Sociability	.797				.203	.766
lavaan package in R						
Extraversion	.879				.121	.863
Assertiveness	.767				.233	.757
Energy level	.674				.326	.647
Sociability	.797				.203	.756
$p^\bullet \times I^\circ \times O^\bullet$ design						
mGENOVA						
Extraversion	.836	.069	.048	.047	.164	.821
Assertiveness	.717	.123	.051	.109	.283	.709
Energy level	.677	.172	.025	.127	.323	.653
Sociability	.771	.129	.036	.064	.229	.729
glmmTBM package in R						
Extraversion	.836	.069	.048	.047	.164	.825
Assertiveness	.717	.124	.051	.109	.283	.710
Energy level	.674	.172	.029	.125	.326	.655
Sociability	.770	.129	.037	.063	.230	.738
lavaan package in R						
Extraversion	.836	.069	.048	.047	.164	.821
Assertiveness	.717	.123	.051	.109	.283	.708
Energy level	.677	.172	.025	.127	.323	.652
Sociability	.771	.129	.036	.064	.229	.729

Note. $n_i' = 4$ per subscale in both designs above, and $n_o' = 1$ in the $p^\bullet \times I^\circ \times O^\bullet$ design. *G* coef = generalizability coefficient; SFE = specific-factor error; TE = transient error; RRE = random-response error; *D* coef = dependability coefficient.

items within a univariate GT *persons* × *splits* × *occasions* design (see Vispoel et al., 2018a, 2018b, 2018c, 2018d; Vispoel, Xu, & Schneider, 2022b), and (c) ignoring division of items into subscales altogether and analyzing all items from the composite as a single domain using univariate GT (see Vispoel, Lee, Chen, & Hong, 2023c; Vispoel et al., 2022, 2023).

In Table 11, we provide composite score *G* coefficients, global *D* coefficients, and measurement error partitioning using the four methods described above for the Extraversion composite score in which $n_i' = 4$ for each subscale and $n_o' = 1$. Because modeling and definitions of domains of generalization vary across these procedures, we would not necessarily expect them to yield the same results. Nevertheless, for the three models in which subscale representation is considered (multivariate GT, bifactor modeling, and parallel splits), *G* coefficients and global *D* coefficients uniformly exceed those for the item-based univariate GT analysis of composite scores in which subscale representation is ignored. Score consistency and variance partitioning indices for the multivariate and bifactor GT designs are virtually identical, but lower in comparison to the *persons* × *splits* × *occasions* design. Score consistency indices are higher for the splits design because measurement error is further reduced by choosing highly correlated splits, equating their means and standard deviations, and balancing subscale representation and directionality of item phrasing across splits (see Vispoel et al., 2018b; Vispoel, Xu, & Schneider, 2022b for guidelines in how to construct parallel splits and computer code in R to aid in their creation). With the multivariate and bifactor GT designs, only subscale representation and directionality of item phrasing are balanced.

Summary and Conclusions

GT provides a comprehensive framework for partitioning observed score variance into multiple sources and using that

Table 10
Partitioning of Variance for $p^\bullet \times i^\circ$ and $p^\bullet \times i^\circ \times o^\bullet$ Designs on Continuous Latent Response Variable Metrics

	Index					
Design/scale	*G* coef/US	SFE	TE	RRE	Total error	Global *D* coef
$p^\bullet \times I^\circ$ design						
Extraversion	.924 (.912, .934)				.076 (.066, .088)	.910 (.894, .921)
Assertiveness	.833 (.804, .857)				.167 (.143, .196)	.823 (.791, .848)
Energy level	.809 (.775, .837)				.191 (.163, .225)	.781 (.741, .812)
Sociability	.876 (.853, .894)				.124 (.106, .147)	.835 (.804, .859)
Mean subscales	.840				.160	.813
Mean all scales	.861				.139	.837
$p^\bullet \times I^\circ \times O^\bullet$ design						
Extraversion	.867 (.846, .886)	.048 (.044, .053)	.061 (.044, .079)	.024 (.021, .026)	.133 (.114, .154)	.853 (.831, .872)
Assertiveness	.778 (.753, .801)	.095 (.086, .106)	.061 (.046, .078)	.065 (.058, .073)	.222 (.199, .247)	.769 (.742, .791)
Energy level	.749 (.722, .774)	.115 (.103, .128)	.079 (.062, .097)	.057 (.050, .064)	.251 (.226, .278)	.725 (.695, .750)
Sociability	.833 (.814, .850)	.092 (.083, .103)	.045 (.036, .055)	.029 (.026, .034)	.167 (.150, .186)	.792 (.768, .811)
Mean subscales	.787	.101	.062	.050	.213	.762
Mean all scales	.807	.088	.062	.044	.193	.785

Note. $n'_i = 4$ per subscale in both designs above, and $n'_o = 1$ for the $p^\bullet \times I^\circ \times O^\bullet$ design. All analyses are based on continuous latent response variable scores corrected for scale coarseness. Values within parentheses for score consistency and measurement error indices in columns 2–7 represent 80% Monte Carlo-based confidence interval limits obtained from the *semTools* package in R (Jorgensen et al., 2022). *G* coefficients are equivalent to proportions of universe score variance. *G* coef = generalizability coefficient; US = universe score; SFE = specific-factor error; TE = transient error; RRE = random-response error; *D* coef = dependability coefficient.

information to improve measurement procedures. Historically, GT has been used predominantly with subjectively scored performance assessments but is equally applicable to objectively scored self-report and other measures. Applications of univariate GT with self-reports have been comprehensively addressed in recent studies for scales considered individually or as part of a profile and underscore the importance of taking multiple sources of measurement error into account when measuring psychological traits (Vispoel et al., 2018a, 2018b, 2018c, 2018d, 2019; Vispoel, Hong, et al., in press; Vispoel, Lee, Chen, & Hong, 2023a; Vispoel & Tao, 2013; Vispoel, Xu, & Kilinc, 2021; Vispoel, Xu, & Schneider, 2022a, 2022b). Our goal here was to demonstrate how multivariate GT can replicate such analyses for individual subscales, while simultaneously extending them to composite scores derived from those subscales. Multivariate GT allows for greater precision in defining universes of generalization, more appropriate indices of score consistency for the global universe(s) of interest, estimates of correlations among subscale scores corrected for relevant sources of measurement error, adjustments for scale coarseness, examination of subscale viability, creation of confidence intervals for key parameters, and mechanisms to determine score consistency for composites when altering weightings and numbers of items and/or occasions for subscale scores. To help readers in applying these techniques, we provide code for running common multivariate GT designs with crossed and nested facets using *mGENOVA* and the *gtheory*, *glmmTMB*, and *lavaan* packages in R. Code and formulas provided in our online supplemental materials are applicable to both objectively scored self-report measures and subjectively scored performance or clinical assessments and encompass all possible combinations of crossed and nested facets for one- and two-facet designs discussed here ($p^\bullet \times i^\circ$, $p^\bullet \times i^\bullet$, $p^\bullet \times i^\circ \times o^\circ$, $p^\bullet \times i^\circ \times o^\bullet$, and $p^\bullet \times i^\bullet \times o^\bullet$). We hope readers find this material useful in applying GT to their own data and extending these frameworks in future research.

Table 11
Partitioning of Observed Score Variance for Alternative GT Designs

	Index								
Design	Initial weights	# of items	# of occasions	*G* coef/US	SFE	TE	RRE	Total error	Global *D* coef
Multivariate $p^\bullet \times I^\circ \times O^\bullet$	.33/.33/.33	4	1	0.836	0.069	0.048	0.047	0.164	0.821
Bifactor $p^\bullet \times I^\circ \times O^\bullet$	.33/.33/.33	4	1	0.837	0.068	0.048	0.048	0.163	0.824
Split-half $p \times S \times O$	.5/.5	6	1	0.859	0.046	0.050	0.045	0.141	0.859
Single-construct $p \times I \times O$	1.00	12	1	0.814	0.091	0.048	0.048	0.186	0.797

Note. *G* coefficients are equivalent to proportions of universe score variance. $n'_o = 1$ in all designs; $n'_i = 4$ per subscale in the multivariate and bifactor designs; $n'_s = 2$ in the split-half design; and $n'_i = 12$ in the single-construct design. *G* coefficient and global *D* coefficients in the bifactor design represent combined general and group factors effects. Within that design, general and group factor effects, respectively, account for proportions of 0.490 and 0.347 of total composite score variance. GT = generalizability theory; *G* coef = generalizability coefficient; US = universe score; SFE = specific-factor error; TE = transient error; RRE = random-response error. *D* coef = dependability coefficient; $p \times S \times O$ = *persons* × *splits* × *occasions*; $p \times I \times O$ = *persons* × *items* × *occasions*.

References

Agresti, A. (2010). *Analysis of ordinal categorical data*. Wiley.

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Waveland Press.

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

Ark, T. K. (2015). *Ordinal generalizability theory using an underlying latent variable framework* [Doctoral dissertation]. University of British Columbia.

Baldwin, P. (2015). Weighting components of a composite score using naïve expert judgments about their relative importance. *Applied Psychological Measurement*, *39*(7), 539–550. https://doi.org/10.1177/0146621615584703

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

Brennan, R. L. (2001b). *Manual for mGENOVA*. Iowa Testing Programs, University of Iowa.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. 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. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Engers, C. K. (2022). *Applied missing data analysis* (2nd ed.). Guilford Press.

Feinberg, R. A., & Wainer, H. (2014). A simple equation to predict a subscore's value. *Educational Measurement: Issues and Practice*, *33*(3), 55–56. https://doi.org/10.1111/emip.12035

Garnier-Villarreal, M., & Jorgensen, T. D. (2020). Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods*, *25*(1), 46–70. https://doi.org/10.1037/met0000224

Geiser, C., & Lockhart, G. (2012). A comparison of four approaches to account for method effects in latent state-trait analyses. *Psychological Methods*, *17*(2), 255–283. https://doi.org/10.1037/a0026977

Grund, S., Robitzsch, A., & Luedtke, O. (2023). *mitml: Tools for multiple imputation in multilevel modeling* (R package version 0.4-4) [Computer software]. https://cran.r-project.org/web/packages/mitml/mitml.pdf

Gulliksen, H. (1950). *Theory of mental tests*. Wiley.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204–229. https://doi.org/10.3102/1076998607302636

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*(2), 209–227. https://doi.org/10.1007/s11336-010-9158-4

Harrell, F. E., Jr., & Dupont, C. (2022). *Hmisc: Harrell miscellaneous* (R package version 4.7-2) [Computer software]. https://cran.r-project.org/web/packages/Hmisc/index.html

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*(7), 1–47. https://doi.org/10.18637/jss.v045.i07

Honaker, J., King, G., & Blackwell, M. (2022). *Amelia: A program for missing data* (R package version 1.8.1) [Computer software]. https://cran.r-project.org/web/packages/Amelia/index.html

Jiang, Z., & Raymond, M. (2018). The use of multivariate generalizability theory to evaluate the quality of subscores. *Applied Psychological Measurement*, *42*(8), 595–612. https://doi.org/10.1177/0146621618758698

Jiang, Z., Raymond, M., Shi, D., & DiStefano, C. (2020). Using a linear mixed-effect model framework to estimate multivariate generalizability theory parameters in R. *Behavior Research Methods*, *52*(6), 2383–2393. https://doi.org/10.3758/s13428-020-01399-z

Jiang, Z., & Skorupski, W. (2018). A Bayesian approach to estimating variance components within a multivariate generalizability theory framework. *Behavior Research Methods*, *50*(6), 2193–2214. https://doi.org/10.3758/s13428-017-0986-3

Jiang, Z., Walker, K., Shi, D., & Cao, J. (2018). Improving generalizability coefficient estimate accuracy: A way to incorporate auxiliary information. *Methodological Innovations*, *11*(2), Article 205979911879139. https://doi.org/10.1177/2059799118791397

Jorgensen, T. D. (2021). How to estimate absolute-error components in structural equation models of generalizability theory. *Psych*, *3*(2), 113–133. https://doi.org/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) [Computer software]. https://CRAN.R-project.org/package=semTools

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160. https://doi.org/10.1007/BF02288391

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. https://doi.org/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. https://doi.org/10.1207/s15328007sem1301_3

LoPilato, A. C., Carter, N. T., & Wang, M. (2015). Updating generalizability theory in management research: Bayesian estimation of variance components. *Journal of Management*, *41*(2), 692–717. https://doi.org/10.1177/0149206314554215

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lumley, T. (2022). *Package 'mitools'* (R package Version 2.4) [Computer software]. https://cran.r-project.org/web/packages/mitools/mitools.pdf

Marcoulides, G. A. (1990). An alternative method for estimating variance components in generalizability theory. *Psychological Reports*, *66*(2), 379–386. https://doi.org/10.2466/pr0.1990.66.2.379

Merkle, E. C. (2011). A comparison of imputation methods for Bayesian factor analysis models. *Journal of Educational and Behavioral Statistics*, *36*(2), 257–276. https://doi.org/10.3102/1076998610375833

Merkle, E. C., & Rosseel, Y. (2018). Blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1–30. https://doi.org/10.18637/jss.v085.i04

Moore, C. T. (2016). *gtheory: Apply generalizability theory with R* (R package version 0.1.2) [Computer software]. https://CRAN.R-project.org/package=gtheory

Muthén, L. K., & Muthén, B. O. (2017). *Mplus: Statistical analysis with latent variables: User's guide (Version 8)*.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, *6*(2), 77–98. https://doi.org/10.1080/19312458.2012.679848

Raymond, M. R., & Jiang, Z. (2020). Indices of subscore utility for individuals and subgroups based on multivariate generalizability theory. *Educational and Psychological Measurement*, *80*(1), 67–90. https://doi.org/10.1177/0013164419846936

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal*

of Personality Assessment, *98*(3), 223–237. https://doi.org/10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

Rosseel, Y. (2012). *lavaan*: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*(2), 199–223. https://doi.org/10.1037/1082-989X.1.2.199

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical investigation of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods*, *8*(2), 206–224. https://doi.org/10.1037/1082-989X.8.2.206

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

Shavelson, R. J., Webb, N. M., & Rowley, G. I. (1992). Generalizability theory. In A. E. Kazdin (Ed.), *Methodological issues & strategies in clinical research* (pp. 233–256). American Psychological Association. https://doi.org/10.1037/10109-051

Sinharay, S. (2019). Added value of subscores and hypothesis testing. *Journal of Educational and Behavioral Statistics*, *44*(1), 25–44. https://doi.org/10.3102/1076998618788862

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Stekhoven, D. J. (2022). *missForest: Nonparametric missing value imputation using random forest* (R package version 1.5) [Computer software]. https://cran.rproject.org/web/packages/missForest/index.html

Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*, *8*(2), 79–98. https://doi.org/10.1027/1015-5759/a000413

Steyer, R., Geiser, C., & Fiege, C. (2012). Latent state-trait models. In H. Copper (Ed.), *Handbook of research methods in psychology: Volume 3. Data analysis and research publication* (pp. 291–308). American Psychological Association. https://doi.org/10.1037/13621-014

Su, Y. S., Yajima, M., Goodrich, B., Si, Y., & Kropko, J. (2022). *mi: Missing data imputation and model checking* (R package version 1.1) [Computer software]. https://cran.r-project.org/web/packages/mi/index.html

van Buuren, S., & Groothuis-Oudshoorn, K. (2022). *mice: Multivariate imputation by chained equations* (R package Version 3.15.0) [Computer software]. https://cran.r-project.org/web/packages/mice/index.html

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. *Structural Equation Modeling*. Advance online publication. https://doi.org/10.1080/10705511.2023.2187734

Vispoel, W. P., Hong, H., Lee, H., & Jorgensen, T. R. (in press). Analyzing complete generalizability theory designs using structural equation models. *Applied Education in Education*.

Vispoel, W. P., & Lee, H. (2023). Merging generalizability theory and bifactor modeling to improve psychological assessments. *Psychology and Psychotherapy: Review Study*, *7*(1), 1–4. https://doi.org/https://crimsonpublishers.com/pprs/pdf/PPRS.000652.pdf

Vispoel, W. P., Lee, H., Chen, T., & Hong, H. (2023a). Using structural equation modeling to reproduce and extend ANOVA-based generalizability theory analyses for psychological assessments. *Psych*, *5*(2), 249–273. https://doi.org/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. https://doi.org/10.3390/psych5020036

Vispoel, W. P., Lee, H., Chen, T., & Hong, H. (2023c). *Analyzing and comparing univariate, multivariate, and bifactor generalizability theory design for hierarchically structured personality traits* [Manuscript submitted for publication].

Vispoel, W. P., Lee, H., & Hong, H. (in press). Analyzing multivariate generalizability theory designs for psychological assessments within structural equation modeling frameworks. *Structural Equation Modeling*.

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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1037/met0000177

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. https://doi.org/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*, *103*(4), 429–442. https://doi.org/10.1080/00223891.2020.1808474

Vispoel, W. P., Xu, G., & Schneider, W. S. (2022a). Interrelationships between latent state-trait theory and generalizability theory within a structural equation modeling framework. *Psychological Methods*, *27*(5), 773–803. https://doi.org/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. https://doi.org/10.1080/00223891.2021.1938589

Received August 18, 2022
Revision received May 17, 2023
Accepted June 22, 2023 ■