Psychological Methods
2018, Vol. 23, No. 1, 1–26

1082-989X/18/$12.00 http://dx.doi.org/10.1037/met0000107

Applications of Generalizability Theory and Their Relations to Classical Test Theory and Structural Equation Modeling

Walter P. Vispoel, Carrie A. Morris, and Murat Kilinc
University of Iowa

Abstract

Although widely recognized as a comprehensive framework for representing score reliability, generalizability theory (G-theory), despite its potential benefits, has been used sparingly in reporting of results for measures of individual differences. In this article, we highlight many valuable ways that G-theory can be used to quantify, evaluate, and improve psychometric properties of scores. Our illustrations encompass assessment of overall reliability, percentages of score variation accounted for by individual sources of measurement error, dependability of cut-scores for decision making, estimation of reliability and dependability for changes made to measurement procedures, disattenuation of validity coefficients for measurement error, and linkages of G-theory with classical test theory and structural equation modeling. We also identify computer packages for performing G-theory analyses, most of which can be obtained free of charge, and describe how they compare with regard to data input requirements, ease of use, complexity of designs supported, and output produced.

Translational Abstract

Generalizability theory (G-theory) is widely recognized as a comprehensive framework for representing score reliability. However, despite its potential benefits, G-theory has been used sparingly in reporting of results for measures of individual differences. In this article, we describe G-theory in a straightforward manner and highlight many valuable ways it can be used to quantify, evaluate, and improve psychometric properties of scores. Our illustrations encompass assessment of overall reliability, percentages of score variation accounted for by individual sources of measurement error, dependability of cut-scores for decision making, estimation of reliability and dependability for changes made to measurement procedures, disattenuation of validity coefficients for measurement error, and linkages of G-theory with classical test theory and structural equation modeling. We also identify computer packages for performing G-theory analyses, most of which can be obtained free of charge, and describe how they compare with regard to data input requirements, ease of use, complexity of designs supported, and output produced. These resources, along with formulas provided throughout the article, should enable readers to apply G-theory to their own research and understand how it aligns with and differs from other measurement models.

Keywords: generalizability theory, reliability, validity, classical test theory, structural equation modeling

Over 40 years have passed since Cronbach, Gleser, Nanda, and Rajaratnam (1972) published their seminal treatise on generalizability theory (G-theory)—*The Dependability of Behavioral Measurements*: *Theory of Generalizability for Scores and Profiles*. Their work significantly broadened perspectives on measurement theory by providing a comprehensive framework for estimating score consistency with reference to multiple sources of measurement error. Over time, many additional treatments of G-theory have appeared that summarize and expand the work of Cronbach et al. (see, e.g., Brennan, 2001a; Crocker & Algina, 1986; Feldt & Brennan, 1989; Haertel, 2006; Marcoulides, 2000; Raykov & Marcoulides, 2011; Shavelson & Webb, 1991; Shavelson, Webb, & Rowley, 1989; Wiley, Webb, & Shavelson, 2013). Yet despite the strong interest in G-theory within the measurement community, applications of it are still rare when reporting results for measures of individual differences. Possible reasons for such neglect may be G-theory's technical vocabulary, overlooked linkages between it and classical test theory (CTT), and difficulty in finding and running software for doing G-theory analyses. The purpose of this article is to describe G-theory in a straightforward manner, illustrate effective ways it can be used with measures of individual differences, highlight many of its direct connections with conventional indices of reliability and validity, show how G-theory can be approached from a structural equation modeling perspective, and identify computer resources for conducting G-theory analyses.

This article was published Online First January 23, 2017.

Walter P. Vispoel, Carrie A. Morris, and Murat Kilinc, Department of Psychological and Quantitative Foundations, University of Iowa.

We thank Patricia Martin for her help in preparing and proof reading drafts of the submitted manuscript.

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. E-mail: walter-vispoel@uiowa.edu

Background

The most important concept in CTT is the reliability of scores for a given measure. The theory begins with the assumption that an individual's observed score (X) is the sum of true (T) and error (E) scores: $X = T + E$. True score represents an individual's expected or average observed score over a presumed infinite number of retakes of the measure with no carryover effects. Because such individual score modeling is impossible to implement in practice, reliability of scores is derived by administering the same measure(s) to a population of individuals. To compute a reliability coefficient for those scores, we assume that parallel forms of a measure can be constructed in which a given individual has the same true score on both forms, observed-score variances are equal across forms, and error scores are uncorrelated with true scores and with each other. Under these conditions, observed-score variance will equal the sum of true-score and error variances, and the correlation between scores from the parallel forms will represent reliability as a ratio of true-score variance over true-score variance plus error variance, as shown in Equation 1:

$$\text{CTT: Reliability coefficient} = \frac{\text{True-score Variance}}{\text{True-score Variance} + \text{Total Error Variance}}. \quad (1)$$

Equation 1 illustrates that error variance in CTT is a single undifferentiated entity. Partitioning of observed-score variance in G-theory can mirror that same partitioning, but refine it further by isolating multiple sources of measurement error. These sources of measurement error in turn define the universe over which results are generalized, leading to the term *universe score* in G-theory replacing the term *true score* in CTT. *Generalizability coefficients* (G-coefficients) in G-theory are analogous to *reliability coefficients* in CTT, and also represent ratios of systematic variance divided by systematic variance plus error variance. The primary difference between the two is that G-coefficients can separate individual sources of measurement error, as shown in Equation 2:

$$\text{G-theory: G-coefficient} = \frac{\text{Universe-score Variance}}{\text{Universe-score Variance} + \text{Individual Sources of Error Variance}}. \quad (2)$$

G-coefficients are typically derived using variance components from analysis of variance (ANOVA) models. In the context of ANOVA, partitioning of true-score and error variances in CTT is conceptually similar to partitioning of systematic and error effects in a one-way design, whereas partitioning of universe score and multiple sources of error in G-theory more closely resembles having multiple effects in a factorial design (Shavelson et al., 1989).

Within a G-theory framework, a single measurement of behavior (item score, subscale score, rating, etc.) is conceptualized as a sample from a universe of admissible observations for the targeted objects of measurement—represented as *persons* in all illustrations discussed here. Aspects of the assessment such as individual items, blocks of items, test forms, prompts, raters, or occasions can represent *facets* or possible sources of measurement error in a G-theory design analogous to *factors* in an ANOVA model. As with factors in an ANOVA model, facets in a G-theory design can be treated as either fixed or random. A facet representing essay prompts, for example, would be considered fixed if inferences are restricted only to the particular prompts administered, or as random if the prompts are viewed as being sampled from a larger universe of possible prompts. In each G-theory design that we illustrate, tasks, occasions, or both will represent measurement facets of interest. Unless otherwise noted, we assume that persons are sampled at random from a target population of interest, and that tasks and occasions are sampled at random from broader universes of similar tasks and occasions, respectively.

In sections to follow, we describe the basics of G-theory with reference to single- and two-facet designs relevant to most measures of individual differences. We then demonstrate using real data how G-theory and conventional reliability coefficients (e.g., alpha, split-half, parallel-form, and test–retest) align and differ. In doing so, we emphasize benefits of two-facet designs in quantifying multiple sources of measurement error and show how those designs can provide more informative indices of overall reliability, dependability of individual cut-scores, and validity coefficients disattenuated for measurement error. In later sections, we describe recent advances in G-theory, its linkages with structural equation modeling, and software packages available for doing G-theory analyses.

G-Theory Basics

Single-Facet Designs

Partitioning of scores. In Table 1, we provide ANOVA models for two single-facet designs. *Task* is the single measurement facet of interest in the first design, and *occasion* in the second. In each design, a given observed score is partitioned into a linear composite representing a grand mean and effects for persons, the measurement facet of interest (task or occasion), and the combination of person and measurement facet. Although our initial illustrations focus on tasks, they can be easily adapted to occasions simply by substituting occasions for tasks in the equations to follow. To show direct linkages between G-theory and CTT indices in a familiar context, we conceptualize tasks in our illustrations as representing items, half-measures (i.e., splits), or full-measures (i.e., forms) for objectively scored instruments such as Likert-style questionnaires or multiple-choice tests in which anyone scoring the measures would get the same results.

Equation 3 represents a G-theory, persons × tasks ($p \times t$) design with *person* as the object of measurement and *task* (item, split, or form) as the measurement facet of interest. This design is a repeated measures, random-effects ANOVA model in which each person has as many scores as number of tasks sampled:

$$\begin{aligned} Y_{pt} &= \mu + (\mu_p - \mu) + (\mu_t - \mu) + (Y_{pt} - \mu_p - \mu_t + \mu) \\ &= \text{grand mean} + \text{person effect} + \text{task effect} + \text{residual}. \end{aligned} \quad (3)$$

In Equation 3, Y_{pt} represents the score for a particular person on a particular task. The grand mean (μ) is a constant that represents the mean Y_{pt} score aggregated across all persons and tasks. The universe score (μ_p), analogous to true score in CTT, corresponds to a person's expected long-run average observed score over the universe of admissible observations included in the model (i.e., tasks in this example). The universe score is typically the primary focus of interest because it is intended to represent a person's score independent of the specific tasks used to derive it. The symbol μ_t

Table 1
G-Theory ANOVA Models, Partitioning, and Score Consistency Indices for One-Facet Designs

Design and Characteristic	Formula
One facet: persons × tasks	
Model[a]	$Y_{pt} = \mu + (\mu_p - \mu) + (\mu_t - \mu) + (Y_{pt} - \mu_p - \mu_t + \mu)$ Score of a person on a task = mean across persons and tasks + person effect + task effect + person × task interaction and other error
Partitioning of variance	Individual score: $\sigma^2_{Y_{pt}} = \sigma^2_p + \sigma^2_t + \sigma^2_{pt,e}$ Mean score: $\sigma^2_{Y_{pT}} = \sigma^2_p + \frac{\sigma^2_{pt,e}}{n'_t}$
Error variances	Relative: $\frac{\sigma^2_{pt,e}}{n'_t}$ Absolute: $\frac{\sigma^2_{pt,e}}{n'_t} + \frac{\sigma^2_t}{n'_t}$
Coefficients	G-coefficient: $\frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_{pt,e}}{n'_t}}$ Global D-coefficient: $\frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_{pt,e}}{n'_t} + \frac{\sigma^2_t}{n'_t}}$
Standard error of measurement	Relative: $\sqrt{\frac{\sigma^2_{pt,e}}{n'_t}}$ Absolute: $\sqrt{\frac{\sigma^2_{pt,e}}{n'_t} + \frac{\sigma^2_t}{n'_t}}$
One facet: persons × occasions	
Model	$Y_{po} = \mu + (\mu_p - \mu) + (\mu_o - \mu) + (Y_{po} - \mu_p - \mu_o + \mu)$ Score of a person on a given occasion = mean across persons and occasions + person effect + occasion effect + person × occasion interaction and other error
Partitioning of variance	Individual score: $\sigma^2_{Y_{po}} = \sigma^2_p + \sigma^2_o + \sigma^2_{po,e}$ Mean score: $\sigma^2_{Y_{pO}} = \sigma^2_p + \frac{\sigma^2_{po,e}}{n'_o}$
Error variances	Relative: $\frac{\sigma^2_{po,e}}{n'_o}$ Absolute: $\frac{\sigma^2_{po,e}}{n'_o} + \frac{\sigma^2_o}{n'_o}$
Coefficients	G-coefficient: $\frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_{po,e}}{n'_o}}$ Global D-coefficient: $\frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_{po,e}}{n'_o} + \frac{\sigma^2_o}{n'_o}}$
Standard error of measurement	Relative: $\sqrt{\frac{\sigma^2_{po,e}}{n'_o}}$ Absolute: $\sqrt{\frac{\sigma^2_{po,e}}{n'_o} + \frac{\sigma^2_o}{n'_o}}$

Note. Primes are used with *n*s in all G-theory based formulas to allow for changes in numbers of replicates within different contexts.
[a] Tasks represent items, splits, or forms in illustrations used throughout this article.

denotes the mean for a particular task aggregated across persons. The deviation score $\mu_p - \mu$ represents the person effect, $\mu_t - \mu$ represents the task effect, and $Y_{pt} - \mu_p - \mu_t + \mu$ is a residual term that reflects what remains in a given Y_{pt} score after the person and task effects are subtracted out. The residual term, often labeled *residual*, *error*, *pt*, or *pt,e*, includes the person × task interaction and any other sources of error.

All components in Equation 3 except the constant μ will have a distribution with a variance representing differences in scores for persons, tasks, or residuals. In this completely crossed, balanced ANOVA model, the variance of Y_{pt} scores can be partitioned into independent additive *variance components* representing persons, tasks, and residuals, as shown in Equation 4:

$$\sigma^2_{Y_{pt}} = \sigma^2_p + \sigma^2_t + \sigma^2_{pt,e}. \quad (4)$$

In the equation, σ^2_p represents the extent to which scores vary across persons, and σ^2_t across tasks; $\sigma^2_{pt,e}$ reflects residual variation in observed scores not accounted for by persons and tasks.

Because observed scores used for decision making in practice usually involve aggregating or averaging across tasks (e.g., summing item scores within a subscale to create a total subscale score), partitioning of those scores is typically of more interest in representing consistency of results. The partitioning of mean scores across tasks is shown in Equation 5. Because means for tasks themselves are constants across individuals, σ^2_t is excluded from that partitioning:

$$\sigma^2_{Y_{pT}} = \sigma^2_p + \frac{\sigma^2_{pt,e}}{n'_t}, \text{ where } n'_t = \text{number of tasks}. \quad (5)$$

Note that the lowercase *t* from Y_{pt} in Equation 4 has been replaced with an uppercase *T* in Equation 5 to reflect the averaging of *Y* scores across all tasks represented.

Indices of score consistency. Indices of score consistency in G-theory are catered to whether scores are used for norm- or criterion-referenced decisions. With norm referencing (e.g., rank ordering), decisions are focused on relative differences in the characteristic of interest. For example, I might want to know where I fall among my peers in extraversion. With criterion referencing, decisions are based on absolute levels of scores. Here, I might be more interested in determining whether my extraversion score is

high enough to qualify me to be hired as a sales representative. For norm-referenced decisions, indices of score consistency will only include variance components that influence relative differences in scores, whereas those for criterion-referenced decisions will include components that reflect both relative and absolute differences, as shown in Equations 6 and 7:

$$\textit{Generalizability}\ \text{(or G-) coefficient} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pt,e}^2}{n_t'}} = \frac{\text{Universe-score Variance}}{\text{Universe-score Variance} + \textit{Relative-error Variance}}; \quad (6)$$

$$\textit{Dependability}\ \text{(or D-) coefficient} = \frac{\sigma_p^2}{\sigma_p^2 + \left[\frac{\sigma_{pt,e}^2}{n_t'} + \frac{\sigma_t^2}{n_t'}\right]} = \frac{\text{Universe-score Variance}}{\text{Universe-score Variance} + \textit{Absolute-error Variance}}. \quad (7)$$

Equations 6 and 7 represent indices of *generalizability* and *dependability* relevant to norm- and criterion-referenced decisions, respectively. In other treatments of G-theory (e.g., Brennan, 2001a), the symbol $E\rho^2$ is often used to denote a G-coefficient, and Φ is used to denote a D-coefficient. In Equation 7, σ_t^2 is included as part of the D-coefficient because the nature of behaviors reflected by tasks could affect the absolute magnitude of scores. Note from Equations 6 and 7 that *absolute-error variance* will always be greater than or equal to *relative-error variance*. They will be equal only when all task means are the same (i.e., $\sigma_t^2 = 0$), thereby reflecting equal levels of the behaviors measured (endorsement, difficulty, etc.). We will refer to D-coefficients like those in Equation 7 as *global* D-coefficients to distinguish them from the *cut-score specific* D-coefficients we consider next.

Global D-coefficients provide overall estimates of consistency accounting for differences in rank order of scores as well as absolute differences in levels of scores. However, in practice, decisions based on absolute levels of scores are typically targeted to specific cut-points. In such cases, cut-score specific dependability indices are of greater interest. Equation 8 represents the general formula for these coefficients within the $p \times T$ design:

$$\text{Cut-score specific D-coefficient} = \frac{\sigma_p^2 + [\mu_Y - C]^2}{\sigma_p^2 + [\mu_Y - C]^2 + \left[\frac{\sigma_{pt,e}^2}{n_t'} + \frac{\sigma_t^2}{n_t'}\right]} = \frac{\text{Universe-score Variance} + [\text{Mean} - \text{Cut-score}]^2}{\text{Universe-score Variance} + [\text{Mean} - \text{Cut-score}]^2 + \text{Absolute-error Variance}} \quad (8)$$

Equation 8 shows that a cut-score specific D-coefficient equals its corresponding global counterpart when the cut-score (C) is at the scale mean (i.e., $\mu_Y - C = 0$), but exceeds that value in other instances. Conceptually, cut-score specific D-coefficients quantify the extent to which an observed score reflects whether an individual is truly above or below the targeted cut-score.

G- and global D-coefficients in G-theory, as well as alpha, split-half, parallel-form, and test–retest coefficients in CTT, provide summary indices of score consistency on a standardized 0 to 1 metric. However, each has the drawback of not being on the scale likely used for decision making (Cronbach, Linn, Brennan, & Haertel, 1997; Cronbach & Shavelson, 2004). In CTT, this drawback can be addressed by transforming a reliability coefficient to the observed score scale using Equation 9 to derive the standard error of measurement (*SEM*), which represents the standard deviation of differences between observed and true scores:

$$\text{SEM}_{\text{CTT}} = \sigma_Y\sqrt{1 - \text{Reliability Coefficient}}. \quad (9)$$

Similarly, standard error indices can be derived from G-theory for making relative and absolute decisions by taking the square roots of relative- and absolute-error variances, as shown in Equations 10 and 11:

$$\text{SEM}_{\text{G-theory, relative}} = \sqrt{\frac{\sigma_{pt,e}^2}{n_t'}} = \sqrt{\text{Relative-error Variance}}; \quad (10)$$

$$\text{SEM}_{\text{G-theory, absolute}} = \sqrt{\frac{\sigma_{pt,e}^2}{n_t'} + \frac{\sigma_t^2}{n_t'}} = \sqrt{\text{Absolute-error Variance}}. \quad (11)$$

If tasks are represented by items or splits, the SEMs from Equations 10 and 11 would need to be multiplied by the number of items (n_i') or splits (n_s') to reference them to the total score metric. Although not discussed here, conditional standard error and associated indices also can be derived for particular points on a score scale using either CTT or G-theory (see, e.g., Brennan, 1998, 2001a; Feldt, 1984; Jarjoura, 1986; Lord, 1957, 1965, 1980; Thorndike, 1951; Vispoel & Tao, 2013; Woodruff, 1991; Woodruff, Traynor, Cui, & Fang, 2013, for further information about how to derive them and the complexities often involved).

Two-Facet Designs

Partitioning of scores. The main problem with reliability indices for the present single-facet designs is that they fail to account for and separate the three primary sources of measurement error typically affecting scores from measures of individual differences: random-response, specific-factor, and transient. Random-response error reflects "noise" affecting scores within a particular occasion of administration resulting from moment-to-moment fluctuations in effort, mood, attention, memory, and other factors. Specific-factor error represents consistent responding to particular tasks unrelated to the construct(s) being measured. Transient error refers to stable factors that affect scores within a particular occasion (fatigue, illness, motivation, etc.) but not across occasions. Unless each source of measurement error is properly taken into account, reliability will likely be overestimated. Reliability indices from single-facet, persons × Tasks ($p \times T$) designs from G-theory and single-occasion alpha, split-half, and parallel-form coefficients from CTT include random-response and specific-factor error, but treat transient error as universe/true-score variance. In contrast, reliability indices for single-facet, persons × Occasions ($p \times O$) designs from G-theory and test–retest coefficients from CTT include random-response and transient error, but treat specific-factor error as universe/true-score variance.

Disentangling these sources of measurement error requires replications of both tasks and occasions that could be represented in a G-theory, persons × tasks × occasions ($p \times t \times o$) design (see Table 2). Within this general design, the task facet again could be items (i), splits (s), or forms (f). As before, the design entails

Table 2
G-Theory ANOVA Models, Partitioning, and Score Consistency Indices for Two-Facet Designs

Characteristic	Formula
Two facet: persons × tasks × occasions	
Model[a]	$Y_{pto} =$ Score of a person on a task on one occasion =
	μ mean across persons, tasks, and occasions
	$+(\mu_p - \mu)$ + person effect (p)
	$+(\mu_t - \mu)$ + task effect (t)
	$+(\mu_o - \mu)$ + occasion effect (o)
	$+(\mu_{to} - \mu_t - \mu_o + \mu)$ + task × occasion interaction ($t \times o$)
	$+(\mu_{pt} - \mu_p - \mu_t + \mu)$ + person × task interaction ($p \times t$)
	$+(\mu_{po} - \mu_p - \mu_o + \mu)$ + person × occasion interaction ($p \times o$)
	$+(Y_{pto} - \mu_{pt} - \mu_{po} - \mu_{to} + \mu_p + \mu_t + \mu_o - \mu)$ + person × task × occasion interaction ($p \times t \times o$) and other error
Partitioning of variance	Individual score: $\sigma^2_{Y_{pto}} = \sigma^2_p + \sigma^2_t + \sigma^2_o + \sigma^2_{to} + \sigma^2_{pt} + \sigma^2_{po} + \sigma^2_{pto,e}$
	Mean score: $\sigma^2_{Y_{pTO}} = \sigma^2_p + \frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pto,e}}{n'_t n'_o}$
Error variances	Relative: $\frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pto,e}}{n'_t n'_o}$ Absolute: $\frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pto,e}}{n'_t n'_o} + \frac{\sigma^2_t}{n'_t} + \frac{\sigma^2_o}{n'_o} + \frac{\sigma^2_{to}}{n'_t n'_o}$
Coefficients	G-coefficient: $\frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pto,e}}{n'_t n'_o}}$ Global D-coefficient: $\frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pto,e}}{n'_t n'_o} + \frac{\sigma^2_t}{n'_t} + \frac{\sigma^2_o}{n'_o} + \frac{\sigma^2_{to}}{n'_t n'_o}}$
Standard error of measurement	Relative: $\sqrt{\frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pto,e}}{n'_t n'_o}}$ Absolute: $\sqrt{\frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pto,e}}{n'_t n'_o} + \frac{\sigma^2_t}{n'_t} + \frac{\sigma^2_o}{n'_o} + \frac{\sigma^2_{to}}{n'_t n'_o}}$

[a] Tasks represent items, splits, or forms in illustrations used throughout this article.

repeated measures but with each person now having as many scores as the product of numbers of tasks and occasions sampled ($n'_t \times n'_o$). Equation 12 shows the partitioning of an observed score Y_{pto} within a $p \times t \times o$, random-effects ANOVA model. Note that a given Y_{pto} score is a linear composite of a grand mean and effects for person, each measurement facet of interest, and all combinations of person and measurement facets:

$$\begin{aligned} Y_{pto} = {} & \mu + (\mu_p - \mu) + (\mu_t - \mu) + (\mu_o - \mu) + (\mu_{to} - \mu_t - \mu_o + \mu) \\ & + (\mu_{pt} - \mu_p - \mu_t + \mu) + (\mu_{po} - \mu_p - \mu_o + \mu) \\ & + (Y_{pto} - \mu_{pt} - \mu_{po} - \mu_{to} + \mu_p + \mu_t + \mu_o - \mu). \end{aligned}$$

Score of a person on a task on one occasion
= mean across persons, tasks, and occasions
+ person effect (p) + task effect (t) + occasion effect (o)
+ task × occasion interaction ($t \times o$)
+ person × task interaction ($p \times t$)
+ person × occasion interaction ($p \times o$)
+ person × task × occasion interaction and other error ($p \times t \times o$, *residual, error*, *pto*, or *pto*, *e*). (12)

In Equation 12, Y_{pto} represents a score for a particular person, task, and occasion. The grand mean (μ) is a constant that equals the mean Y score aggregated across all persons, tasks, and occasions. The universe score (μ_p) represents a person's expected long-run average observed score over all combinations of tasks and occasions. The symbol μ_t is the mean for a particular task aggregated across persons and occasions; μ_o is the mean for a particular occasion aggregated across persons and tasks; and $\mu_p - \mu$, $\mu_t - \mu$, and $\mu_o - \mu$, represent main effects for person, task, and occasion, respectively. The remaining components in the model reflect all possible two- and three-way interactions involving persons and the measurement facets of interest.

A task × occasion ($t \times o$) interaction effect would indicate that differences in task means vary by occasion. A person × task ($p \times t$) interaction effect would reveal that differences in task means vary from person to person. These idiosyncratic task differences in scores signal the presence of specific-factor error. A person × occasion ($p \times o$) interaction effect would indicate that differences in occasion means vary from person to person. These person-specific occasion differences reflect the presence of transient error. The person × task × occasion interaction ($p \times t \times o$) represents what remains after all other main and interaction effects are subtracted from Y_{pto}. This term, typically labeled as *residual*, *error*, *pto*, or *pto,e*, is treated as random-response error and includes the three-way interaction and other sources of error unaccounted for in the model.

As was the case with the single-facet designs, the variance of individual scores in this two-facet design will be a composite of variance components for all main and interaction effects, as shown in Equation 13:

$$\sigma^2_{Y_{pto}} = \sigma^2_p + \sigma^2_t + \sigma^2_o + \sigma^2_{to} + \sigma^2_{pt} + \sigma^2_{po} + \sigma^2_{pto,e}. \quad (13)$$

However, partitioning again is usually more meaningful when individual behavioral scores are aggregated or averaged across

tasks, occasions, or both. The partitioning for scores averaged across both tasks and occasions is shown in Equation 14:

$$\sigma^2_{Y_{pTO}} = \sigma^2_p + \frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pto,e}}{n'_t n'_o}. \quad (14)$$

Indices of score consistency. Equations 15 and 16 represent the G- and D-coefficients for the $p \times T \times O$ design:

$$\text{G-coefficient} = \frac{\sigma^2_p}{\sigma^2_p + \left[\frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pto,e}}{n'_t n'_o}\right]} = \frac{\text{Universe-score Variance}}{\text{Universe-score Variance} + \text{Relative-error Variance}}; \quad (15)$$

$$\text{D-coefficient} = \frac{\sigma^2_p}{\sigma^2_p + \left[\frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pto,e}}{n'_t n'_o} + \frac{\sigma^2_t}{n'_t} + \frac{\sigma^2_o}{n'_o} + \frac{\sigma^2_{to}}{n'_t n'_o}\right]} = \frac{\text{Universe-score Variance}}{\text{Universe-score Variance} + \text{Absolute-error Variance}}. \quad (16)$$

A crucial difference between representations of relative error in these equations compared with Equations 6 and 7 for the single-facet designs is that three sources of measurement error variance are separately represented, with $\frac{\sigma^2_{pt}}{n'_t}$ equaling specific-factor error, $\frac{\sigma^2_{po}}{n'_o}$ equaling transient error, and $\frac{\sigma^2_{pto,e}}{n'_t n'_o}$ equaling random-response error. Effects for tasks, occasions, and their interaction are included in the denominator for the D-coefficient but not the G-coefficient because those effects can change the absolute magnitude of scores but not their relative differences. Standard errors of measurement again can be derived by taking the square roots of relative- and absolute-error variance, as shown in Equations 17 and 18:

$$\text{SEM}_{\text{G-theory, relative}} = \sqrt{\frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pto,e}}{n'_t n'_o}} = \sqrt{\text{Relative-Error Variance}}; \quad (17)$$

$$\text{SEM}_{\text{G-theory, absolute}} = \sqrt{\frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pto,e}}{n'_t n'_o} + \frac{\sigma^2_t}{n'_t} + \frac{\sigma^2_o}{n'_o} + \frac{\sigma^2_{to}}{n'_t n'_o}} = \sqrt{\text{Absolute-error Variance}}. \quad (18)$$

As before, these indices need to be multiplied by number of items (n'_i) in the $p \times I \times O$ design and number of splits (n'_s) in the $p \times S \times O$ design to put them on the total score metric.

Coefficients of score consistency for two-facet designs are highly desirable because they can separate specific-factor, transient, and random-response measurement error. The G-coefficient for this design shown in Equation 15 is sometimes called a *coefficient of equivalence and stability* (CES) because it allows for generalization over both tasks and occasions. However, variance components from this design also can be used to derive a G-coefficient reflecting generalization only over tasks for fixed occasions (i.e., a *coefficient of equivalence* [CE]), or only over occasions for fixed tasks (i.e., a *coefficient of stability* [CS]). These coefficients are conceptually parallel to G-coefficients from the single-facet $p \times T$ and $p \times O$ designs discussed earlier.

Formulas for computing CEs and CSs from the $p \times T \times O$ design are given in Equations 19 and 20:

G-coefficient of equivalence (CE) for the $p \times T \times O$ design.

$$= \frac{\sigma^2_p + \frac{\sigma^2_{po}}{n'_o}}{\sigma^2_p + \frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pto,e}}{n'_t n'_o}} = \frac{\text{Universe-score} + \text{Transient-error Variance}}{\text{Universe-score} + \text{Specific-factor-} + \text{Transient-} + \text{Random-response-error Variance}}; \quad (19)$$

G-coefficient of stability (CS) for the $p \times T \times O$ design.

$$= \frac{\sigma^2_p + \frac{\sigma^2_{pt}}{n'_t}}{\sigma^2_p + \frac{\sigma^2_{pt}}{n'_t} + \frac{\sigma^2_{po}}{n'_o} + \frac{\sigma^2_{pto,e}}{n'_t n'_o}} = \frac{\text{Universe-score} + \text{Specific-factor-error Variance}}{\text{Universe-score} + \text{Specific-factor-} + \text{Transient-} + \text{Random-response-error Variance}} \quad (20)$$

Equations 19 and 20 demonstrate that CEs analogous to single-occasion alpha, split-half, and parallel-form coefficients treat transient error as universe/true-score variance, whereas CSs analogous to test–retest coefficients treat specific-factor error as universe/true-score variance. Occasion is a *hidden facet* in a single-facet $p \times T$ design, and task is a *hidden facet* in a single-facet $p \times O$ design, whose effects cannot be separated from universe/true-score variance. In the numerators of Equations 19 and 20, this confounding is made explicit. If the intent is to generalize over both tasks and occasions, then CESs computed from Equation 15 clearly are more appropriate indices to report in practice than are either CEs or CSs. A key advantage of G-theory over CTT reliability coefficients is that universes of generalization are unambiguously defined.

Hidden facets representing excluded sources of measurement error are present in even very complex G-theory designs (Cronbach et al., 1997; Haertel, 2006). For example, a possible hidden facet in the $p \times T \times O$ design might be task ordering. One way to address such effects would be to include sets of task orderings as part of the design. This would allow for the variance associated with ordering effects to be accounted for and explicitly estimated. Alternatively, not varying the ordering of tasks would lead to such effects contributing to both the numerator and denominator of G- and D-coefficients. To reduce possible effects of hidden facets on indices of score consistency in practice, we would try to include as many relevant sources of measurement error in the design as is practical.

In the next section, we use real data from a variety of Likert-style, self-report measures to illustrate applications of G-theory and highlight their parallels in CTT. We then discuss how G-theory can be used to optimize a measurement procedure and disattenuate correlation coefficients for measurement error.

G-Theory Analyses

Data Source and Descriptive Statistics

Our illustrations of G-theory are based on data obtained from 206 college students who completed web-based versions of the Big Five Inventory (BFI; John, Donahue, & Kentle, 1991), Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1991), and parallel forms of the Texas Social Behavior Inventory (TSBI; Helmreich & Stapp, 1974). Students completed these questionnaires on two occasions with a 1-week interval between administrations.

The BFI has 44 items scored on a 5-point scale (1 = *disagree strongly*, 5 = *agree strongly*) that measure five broad dimensions of personality. One 10-item scale measures Openness, two nine-item scales measure Agreeableness and Conscientiousness, and two eight-item scales measure Extraversion and Neuroticism. The BIDR has 40 items that measure two dimensions of socially desirable responding: Impression Management (IM) and Self-Deceptive Enhancement (SDE). Each dimension is assessed using a 20-item subscale scored on a 7-point metric (1 = *not true*, 7 = *very true*). The TSBI is described by its authors as an "objective measure of self-esteem or social competence" (Helmreich & Stapp, 1974, p. 473). They created two 16-item parallel forms from an original 32-item version of the instrument with items scored on a 5-point scale (1 = *not at all characteristic of me*, 5 = *very much characteristic of me*). For all scales across instruments, we reverse scored negatively keyed items and summed item scores to derive total subscale scores. Descriptive statistics for subscale scores for all instruments on both occasions appear in Table 3.

We provide conventional (alpha, split-half, parallel-form, test–retest) and G-theory based reliability coefficients for all subscales in Table 4. Split-half coefficients were computed using Rulon's (1939) formula for splits with the same numbers of items, and Raju's (1977) formula for splits with different numbers of items as shown in Table A1 of Appendix A. With even splits, both formulas yield the same result. We also report split-half reliability coefficients for even splits using the Spearman-Brown formula for comparative purposes (see Equation 37 appearing later in the text).

Table 3
Descriptive Statistics for BFI, BIDR, and TSBI Subscale Scores

Scale	Number of items	Occasion 1		Occasion 2	
		M	*SD*	*M*	*SD*
BFI					
Agreeableness	9	35.85	5.05	36.11	5.35
Conscientiousness	9	33.92	5.40	33.85	5.53
Extraversion	8	27.20	6.15	27.32	6.06
Neuroticism	8	24.38	6.30	24.10	6.35
Openness	10	35.69	5.67	35.65	6.51
BIDR					
Impression Management	20	79.80	16.60	82.13	18.40
Self-Deceptive Enhancement	20	82.71	13.56	85.44	13.99
TSBI					
Form A	16	39.70	9.11	39.54	10.07
Form B	16	41.86	9.52	41.52	10.55

Note. BFI = Big Five Inventory; BIDR = Balanced Inventory of Desirable Responding; TSBI = Texas Social Behavior Inventory.

Splits were created using data from previous samples of similar respondents based on balancing of positively and negatively phrased items, similarity of split means and standard deviations, and high intersplit correlations. We applied the same criteria to uneven splits except that means and standard deviations were balanced at the item-mean level.

G-Coefficients and Their Conventional Counterparts

Single-facet designs. Variance components, estimated using the software package urGENOVA (Brennan, 2001c), are provided in Table 5 for all single-facet designs. The largest in magnitude are for persons (reflecting differences in universe scores among respondents), items (reflecting differences in item means within a subscale), uneven splits (reflecting artifactual differences in means for uneven splits), and the person × task or person × occasion interactions (reflecting the presence of measurement error). The smallest variance components are generally for occasion and even splits, supporting stability of mean scale scores over the 1-week time interval between administrations and good matching of splits when using the same number of items.

In Table A1 of Appendix A, we provide formulas for computing the G-coefficient for each design and its corresponding conventional counterpart. The formulas for G-coefficients for the single-facet designs are the same as those presented earlier in Table 1 except that variances are estimates of population parameters. Computation of the estimated G-coefficient for the $p \times I$ design for the BFI Extraversion scale on Occasion 1 is illustrated in Equation 21:

$$\text{Estimated G-coefficient for BFI Extraversion}_{p\times I} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi,e}^2}{n_i'}} = \frac{.505}{.505 + \frac{.695}{8}} = .853. \quad (21)$$

G-theory and CTT reliability coefficients shown in Table 4 are reported to three decimal places to compare their relative magnitudes more precisely. Inspection of the table reveals that G-coefficients for the $p \times I$ design are identical to corresponding alpha coefficients, and those for the $p \times S$ design are identical to Rulon split-half coefficients. Although initially differing, G-coefficients for splits with unequal numbers of items can be made equal to Raju split-half coefficients by multiplying the G-coefficient by $0.25 \times \left(\frac{\text{total number of items}}{\text{number of items in split 1}} \times \frac{\text{total number of items}}{\text{number of items in split 2}}\right)$. This is the same way that a conventional Rulon split-half coefficient computed using uneven splits is transformed to a Raju split-half coefficient. Note that after such adjustments are made (see the column labeled $p \times S$ (adj.) in Table 4), G-coefficients and Raju split-half coefficients are identical. Although generally very close in value, Spearman-Brown coefficients, which are applicable only to even splits, are greater than or equal to corresponding Rulon coefficients. They would be equal to each other only when splits have equal variances (Crocker & Algina, 1986; Cronbach, 1951).

The values of G-coefficients for the $p \times F$ designs in Table 4 are also very close to but not always the same as their conventional counterparts, correlations between parallel forms. The discrepancy results from differences in denominators used to calculate the

Table 4
Conventional and G-Theory Reliability Coefficients Associated With Single-Facet Designs

	Equivalence									Stability	
	Conventional					G-theory					
Scale	Alpha	Rulon	Raju	SB	Forms correlation	$p \times I$	$p \times S$	$p \times S$ (adj.)	$p \times F$	Conventional test–retest	G-theory $p \times O$
					Occasion 1						
BFI											
Agreeableness	.756	.798	.808			.756	.798	.808		.804	.802
Conscientiousness	.793	.853	.864			.793	.853	.864		.832	.832
Extraversion	.853	.877		.877		.853	.877			.896	.896
Neuroticism	.837	.875		.875		.837	.875			.864	.864
Openness	.774	.804		.804		.774	.804			.876	.868
BIDR											
IM	.781	.818		.818		.781	.818			.876	.871
SDE	.711	.754		.755		.711	.754			.807	.807
TSBI											
Form A	.837	.879		.879	.871	.837	.879		.870	.840	.836
Form B	.864	.920		.921		.864	.920			.831	.827
					Occasion 2						
BFI											
Agreeableness	.812	.853	.864			.812	.853	.864			
Conscientiousness	.814	.848	.859			.814	.848	.859			
Extraversion	.863	.894		.894		.863	.894				
Neuroticism	.849	.890		.890		.849	.890				
Openness	.831	.877		.878		.831	.877				
BIDR											
IM	.836	.896		.896		.836	.896				
SDE	.754	.804		.804		.754	.804				
TSBI											
Form A	.871	.911		.911	.899	.871	.911		.898		
Form B	.885	.927		.927		.885	.927				

Note. SB = Spearman-Brown; adj. = adjusted; BIDR = Balanced Inventory of Desirable Responding; IM = Impression Management; SDE = Self-Deceptive Enhancement; TSBI = Texas Social Behavior Inventory; BFI = Big Five Inventory.

coefficients, reflecting their contrasting roots in ANOVA and correlation. The denominator for the G-coefficient represents an *arithmetic mean*—an average of the observed score variances for the two forms ($(\hat{\sigma}^2_{Form_1} + \hat{\sigma}^2_{Form_2})/2$), whereas the denominator for the parallel-form coefficient represents a corresponding *geometric mean* ($\sqrt{\hat{\sigma}^2_{Form_1} \times \hat{\sigma}^2_{Form_2}}$; see Table A1). Because a geometric mean is less than or equal to an arithmetic mean, a parallel-form coefficient will be greater than its corresponding G-coefficient unless score variances for the two forms are equal. The same can be said about relations between the G-coefficients for the $p \times O$ design and their conventional counterparts, test–retest coefficients. These coefficients will be identical only when occasion variances are equal, and greater for test–retest coefficients in other instances. Parallel-form and test–retest coefficients can be converted to their G-theory counterparts by multiplying them by the ratio of the geometric mean over the arithmetic mean or by dividing the numerator of the conventional coefficient by the arithmetic rather than geometric mean of the variances represented in its denominator (see Table A1).

Relations between single-facet G-coefficients and CTT reliability coefficients are summarized in Equations 22 through 26:

$$\text{Items: G-Coefficient}_{p\times I\text{ Design}} = \text{Coefficient Alpha;} \quad (22)$$

Even Splits:

$$\text{G-Coefficient}_{p\times S\text{ Design}} = \text{Rulon}_{SH} \leq \text{Spearman-Brown}_{SH}; \quad (23)$$

Uneven Splits:

$$0.25 \times \left(\frac{\text{total number of items}}{\text{number of items in split 1}} \times \frac{\text{total number of items}}{\text{number of items in split 2}}\right) \times \text{G-Coefficient}_{p\times S\text{ Design}} = \text{Raju}_{SH}; \quad (24)$$

$$\text{Forms: G-Coefficient}_{p\times F\text{ Design}} \leq \text{CTT Parallel Forms;} \quad (25)$$

$$\text{Occasions: G-Coefficient}_{p\times O\text{ Design}} \leq \text{CTT Test-Retest.} \quad (26)$$

Equations 22, 23, and 24 show that G-coefficients coincide with alpha and Rulon coefficients and will equal Raju coefficients if multiplied by a constant. Equations 23, 25, and 26 reiterate that G-coefficients will be less than or equal to corresponding Spearman-Brown split-half, parallel-form, and test–retest coefficients. They would coincide only when scores for splits, forms, or occasions have the same variance, and this would occur whenever scores are classically parallel.

Relationships between standard errors of measurement for G-theory and CTT will mirror those for reliability coefficients, but in the opposite direction. Higher reliability coefficients will yield lower standard errors, and vice versa. Once either a G-theory or CTT relative reliability coefficient is derived, it can be referenced to the score scale of interest using Equation 27.

$$\text{SEM}_{\text{score scale}} = \hat{\sigma}_{\text{score scale}}\sqrt{1 - \text{Relative Reliability Estimate}}. \quad (27)$$

Table 5
Variance Components for Single-Facet Designs

Design	Effect	BFI					BIDR		TSBI	
		Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness	IM	SDE	Form A	Form B
					Occasion 1					
$p \times i$	person (p)	.238	.286	.505	.519	.249	.538	.327	.271	.306
	item (i)	.157	.388	.295	.336	.323	1.036	.412	.092	.198
	$p \times i$, *error*	.690	.671	.695	.810	.725	3.016	2.655	.847	.772
$p \times s$	person (p)	5.087	6.228	8.302	8.689	6.455	56.314	34.678	18.212	20.845
	splits (s)	5.860	6.098	.279	.123	.480	−.119	−.107	−.024	.024
	$p \times s$, *error*	2.577	2.145	2.335	2.477	3.144	25.136	22.580	5.029	3.605
$p \times o$	person (p)	21.688	24.871	33.425	34.582	32.346	267.523	153.179	77.062	83.415
	occasion (o)	.006	−.022	−.012	.015	−.023	2.512	3.530	−.061	−.027
	$p \times o$, *error*	5.344	5.026	3.866	5.430	4.923	39.491	36.622	15.129	17.488
$p \times f$	person (p)								75.499	
	form (f)								2.279	
	$p \times f$, *error*								11.248	
					Occasion 2					
$p \times i$	person (p)	.286	.308	.495	.535	.353	.708	.369	.345	.384
	item (i)	.116	.376	.230	.260	.246	.904	.318	.109	.160
	$p \times i$, *error*	.598	.632	.628	.762	.717	2.776	2.405	.819	.801
$p \times s$	person (p)	6.090	6.487	8.204	8.968	9.305	75.805	39.330	23.105	25.766
	splits (s)	6.098	6.581	.106	.085	.180	−.059	−.074	−.010	.019
	$p \times s$, *error*	2.101	2.322	1.945	2.220	2.606	17.640	19.205	4.529	4.076
$p \times f$	person (p)								95.492	
	form (f)								1.899	
	$p \times f$, *error*								10.856	

Note. BFI = Big Five Inventory; BIDR = Balanced Inventory of Desirable Responding; TSBI = Texas Social Behavior Inventory; IM = Impression Management; SDE = Self-Deceptive Enhancement.

Two-facet designs. Variance components for all two-facet designs and subscales appear in Table 6. The designs include main effects for persons, tasks (i.e., items, splits, or forms), and occasions, and all corresponding interactions. The largest variance components are for persons (reflecting differences in universe scores among respondents), items (reflecting differences in item means within a subscale), uneven splits (reflecting artifactual differences in means between splits with different numbers of items), and the person × task, person × occasion, and person × task × occasion interactions (reflecting, respectively, specific-factor, transient, and random-response measurement error). The smallest components overall are for even splits, occasions, and task × occasion interactions. In Table A2 of Appendix A, we provide formulas for computing G-coefficients for each design and their conventional counterparts.

Formulas in Tables A1 and A2 reveal that conventional CSs for the $p \times O$, $p \times I \times O$, and $p \times S \times O$ designs all correspond to the CTT test–retest coefficient. Although not obvious from their formulas, G-theory CSs associated with these designs also will yield a common result but one not necessarily equivalent to the CTT test–retest coefficient resulting from use of arithmetic rather than geometric means in the formulas. Congruence in the G-theory CSs occurs because the same tasks are treated as fixed in all three designs. Conventional counterparts to the G-theory CEs for items, splits, and forms in Table A2 were created by combining the covariance terms for both occasions in the numerator and dividing by the geometric mean of the corresponding occasion variances in the denominator. Conventional formulas for CESs shown in the table were taken from Green (2003) for items, and from Schmidt, Le, and Ilies (2003) for splits. We generalized their approaches to create an analogous formula for forms.

In Table 7, we provide G-theory-based, two-facet design CEs, CSs, and CESs for all subscales and their corresponding conventional counterparts computed using the formulas given in Table A2. Estimation of G-coefficients for the BFI Openness scale from the $p \times I \times O$ design is demonstrated in Equations 28 to 30. Analogous standard errors of measurement can be computed using Equation 27.

$$\text{CE for BFI Openness}_{p\times I\times O} = \frac{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{po}^2}{n'_o}}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi}^2}{n'_i} + \frac{\hat{\sigma}_{po}^2}{n'_o} + \frac{\hat{\sigma}_{pio,e}^2}{n'_i n'_o}}$$

$$= \frac{.282 + \frac{.019}{1}}{.282 + \frac{.417}{10} + \frac{.019}{1} + \frac{.304}{10\times 1}} = .807; \quad (28)$$

$$\text{CS for BFI Openness}_{p\times I\times O} = \frac{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi}^2}{n'_i}}{\sigma_p^2 + \frac{\hat{\sigma}_{pi}^2}{n'_i} + \frac{\hat{\sigma}_{po}^2}{n'_o} + \frac{\hat{\sigma}_{pio,e}^2}{n'_i n'_o}}$$

$$= \frac{.282 + \frac{.417}{10}}{.282 + \frac{.417}{10} + \frac{.019}{1} + \frac{.304}{10\times 1}} = .868; \quad (29)$$

Table 6
Variance Components for Two-Facet Designs

	BFI					BIDR		TSBI	
Design and Effect	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness	IM	SDE	Form A	Form B
$p \times i \times o$									
person (p)	.235	.270	.475	.489	.282	.588	.313	.275	.301
item (i)	.136	.383	.261	.297	.281	.964	.352	.099	.178
occasion (o)	.000	.000	−.001	.000	−.001	.006	.008	.000	.000
$i \times o$	.000	−.001	.002	.001	.004	.006	.013	.002	.001
$p \times i$	.298	.338	.375	.414	.417	1.610	1.393	.423	.395
$p \times o$	.028	.027	.025	.038	.019	.035	.035	.034	.044
$p \times i \times o$, *error*	.346	.313	.286	.372	.304	1.284	1.137	.410	.392
$p \times s \times o$									
person (p)	4.979	5.737	7.937	8.150	7.326	61.728	33.066	18.161	20.086
split (s)	5.985	6.341	.179	.109	.302	−.060	−.058	−.011	.032
occasion (o)	.004	−.005	−.010	.006	−.020	.642	.899	−.012	−.001
$s \times o$	−.006	−.002	.014	−.005	.028	−.029	−.033	−.006	−.011
$p \times s$	.886	.982	.838	.991	1.521	10.305	10.458	2.210	1.536
$p \times o$	.610	.631	.316	.678	.554	4.331	3.938	2.498	3.220
$p \times s \times o$, *error*	1.453	1.252	1.302	1.358	1.354	11.084	10.435	2.570	2.305
$p \times f \times o$									
person (p)								77.868	
form (f)								2.123	
occasion (o)								−.011	
$f \times o$								−.034	
$p \times f$								2.370	
$p \times o$								7.627	
$p \times f \times o$, *error*								8.682	

Note. BFI = Big Five Inventory; BIDR = Balanced Inventory of Desirable Responding; TSBI = Texas Social Behavior Inventory; IM = Impression Management; SDE = Self-Deceptive Enhancement.

$$\text{CES for BFI Openness}_{p\times I\times O} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi}^2}{n_i'} + \frac{\hat{\sigma}_{po}^2}{n_o'} + \frac{\hat{\sigma}_{pio,e}^2}{n_i' n_o'}}$$

$$= \frac{.282}{.282 + \frac{.417}{10} + \frac{.019}{1} + \frac{.304}{10 \times 1}} = .756. \quad (30)$$

In all cases, G-coefficients for the two-facet designs in Table 7 are less than or equal to corresponding conventional coefficients, reflecting use of arithmetic means in the denominators for G-coefficients and geometric means for conventional coefficients (see Table A2). As with the single-facet designs, G-theory and conventional coefficients for the two-facet designs will be the same if variances for tasks and/or occasions are equal. Similarly, conventional coefficients can be converted to G-coefficients by multiplying them by constants involving ratios of geometric over arithmetic means shown in the footnote to Table A2, or by dividing their numerators by the arithmetic rather than geometric mean of the relevant variances. For either G-theory or conventional indices, CEs and CESs in Table 7 are greater for splits and forms than for items, reflecting closer similarity in variances for larger blocks of tasks.

To probe further into differences among reliability estimates, we provide percentages of universe/true-score, specific-factor error, transient error, and random-response error variance for all two-facet designs and subscales in Table 8. Disentangling these separate sources of measurement error facilitates comparison of their relative effects on scores and provides guidance for changing a measurement procedure to enhance reliability. Large proportions of specific-factor error might be reduced by lengthening a measure, and large proportions of transient error by pooling results across two or more occasions.

For both G-theory and CTT analyses, once CEs, CSs, and CESs are derived, percentages for universe/true score can be computed by multiplying CES by 100, for specific-factor error by multiplying (CS − CES) by 100, for transient error by multiplying (CE − CES) by 100, and for random-response error by multiplying (1 − CE − CS + CES) by 100. As was the case for CEs, CSs, and CESs, results in Table 8 for percentages of score variation show a close correspondence across G-theory and conventional analyses. The primary difference in results overall is that specific-factor-error variance is greater for items than for splits or forms, and this in turn leads to lower corresponding estimates of CEs and CESs. Such differences would be expected when combining items to create comparable splits and forms. Across all G-theory and conventional analyses, CESs are overestimated by 3.98% to 16.03% using CEs (*Mdn*s = 10.11%, 11.02%, and 9.79% for items, splits, and forms, respectively) and by 3.04% to 22.23% using CSs (*Mdn*s = 13.70%, 7.23%, and 3.04% for items, splits, and forms, respectively). These results convincingly underscore problems with relying solely on either CEs or CSs in practice and the usefulness of CESs in generally providing more appropriate and comprehensive indices of score consistency.

D-Coefficients and Their Conventional Counterparts

Formulas for computing estimates of both global and cut-score specific D-coefficients for all single- and two-facet designs are provided in Table A3 of Appendix A. Additional guidelines for deriving more accurate D-coefficients when using splits with un-

Table 7
Conventional and G-Theory Reliability Coefficients for Two-Facet Designs

Index	BFI					BIDR		TSBI	
	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness	IM	SDE	Form A	Form B
Equivalence									
Conventional									
CE Items-based	.787	.804	.858	.843	.814	.816	.734	.860	.880
CE Splits-based	.828	.851	.885	.883	.854	.865	.780	.901	.929
CE Splits-based (adj.)	.839	.862							
CE Forms-based								.891	
Two-facet G-theory									
CE $p \times I \times O$	.786	.804	.858	.843	.807	.811	.733	.855	.875
CE $p \times S \times O$	.827	.851	.885	.883	.846	.861	.780	.896	.924
CE $p \times S \times O$ (adj.)	.837	.861							
CE $p \times F \times O$								.886	
Stability									
Conventional									
Test–retest	.804	.832	.896	.864	.876	.876	.807	.840	.831
CS Forms-based								.836	
Two-facet G-theory									
CS $p \times I \times O$	.802	.832	.896	.864	.868	.871	.807	.836	.827
CS $p \times S \times O$	.802	.832	.896	.864	.868	.871	.807	.836	.827
CS $p \times F \times O$								.831	
Equivalence and Stability									
Conventional									
CES Items-based	.704	.730	.816	.782	.763	.771	.661	.766	.768
CES Splits-based	.738	.766	.852	.815	.794	.809	.697	.792	.800
CES Splits-based (adj.)	.747	.776							
CES Forms-based								.812	
Two-facet G-Theory									
CES $p \times I \times O$	.703	.730	.816	.782	.756	.766	.660	.763	.764
CES $p \times S \times O$	.737	.766	.851	.815	.786	.804	.697	.788	.796
CES $p \times S \times O$ (adj.)	.746	.776							
CES $p \times F \times O$								.807	

Note. BFI = Big Five Inventory; BIDR = Balanced Inventory of Desirable Responding; TSBI = Texas Social Behavior Inventory; IM = Impression Management; SDE = Self-Deceptive Enhancement; CE = Coefficient of Equivalence; adj. = adjusted; CS = Coefficient of Stability; CES = Coefficient of Equivalence and Stability.

equal numbers of items are given in Appendix B. Computation of the Global D-coefficient for the BIDR IM scale from the $p \times I \times O$ design is shown in Equation 31:

$$\text{Estimated Global D-coefficient for } IM_{p\times I\times O} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \left(\frac{\hat{\sigma}_{pi}^2}{n_i'} + \frac{\hat{\sigma}_{po}^2}{n_o'} + \frac{\hat{\sigma}_{pio,e}^2}{n_i'n_o'} + \frac{\hat{\sigma}_i^2}{n_i'} + \frac{\hat{\sigma}_o^2}{n_o'} + \frac{\hat{\sigma}_{io}^2}{n_i'n_o'}\right)}$$
$$= \frac{.588}{.588 + \left(\frac{1.610}{20} + \frac{.035}{1} + \frac{1.284}{20\times 1} + \frac{.964}{20} + \frac{.006}{1} + \frac{.006}{20\times 1}\right)}$$
$$= .715. \quad (31)$$

As mentioned earlier, decisions based on absolute levels of scores are typically targeted to specific cut-points on the score scale. In such cases, cut-score specific D-coefficients are of greater interest than global coefficients. The formula for estimating cut-score specific D-coefficients for the $p \times T \times O$ design is shown in Equation 32. Note that a new term ($\hat{\sigma}_{\bar{Y}}^2$), representing a correction for bias, appears in the numerator and denominator (see Brennan & Kane, 1977):

$$\text{Estimated Cut-score specific D-coefficient} = \frac{\hat{\sigma}_p^2 + \left[(\bar{Y} - C)^2 - \hat{\sigma}_{\bar{Y}}^2\right]}{\hat{\sigma}_p^2 + \left[(\bar{Y} - C)^2 - \hat{\sigma}_{\bar{Y}}^2\right] + \left(\frac{\hat{\sigma}_{pt}^2}{n_t'} + \frac{\hat{\sigma}_{po}^2}{n_o'} + \frac{\hat{\sigma}_{pto,e}^2}{n_t'n_o'} + \frac{\hat{\sigma}_t^2}{n_t'} + \frac{\hat{\sigma}_o^2}{n_o'} + \frac{\hat{\sigma}_{to}^2}{n_t'n_o'}\right)}, \quad (32)$$

where $\hat{\sigma}^2$ = estimated variance, $\bar{Y}$ = grandmean on task scale, C = value of cut-score on task scale, p = person, t = task (item, split, or form), o = occasion, to = task × occasion interaction, pt = person × task interaction, po = person × occasion interaction, pto,e = person × task × occasion interaction and other error, $\hat{\sigma}_{\bar{Y}}^2 = \frac{\hat{\sigma}_p^2}{n_p'} + \frac{\hat{\sigma}_{pt}^2}{n_p'n_t'} + \frac{\hat{\sigma}_{po}^2}{n_p'n_o'} + \frac{\hat{\sigma}_{pto,e}^2}{n_p'n_t'n_o'} + \frac{\hat{\sigma}_t^2}{n_t'} + \frac{\hat{\sigma}_o^2}{n_o'} + \frac{\hat{\sigma}_{to}^2}{n_t'n_o'}$, n_p' = number of persons, n_o' = number of occasions, and n_t' = number of tasks.

A good example of applying cut-scores in decision making is using the BIDR IM subscale to detect possible faked responses. Items on the IM scale reflect desirable but uncommon behaviors, with markedly high and low endorsement of those behaviors possibly signaling attempts to fake good and fake bad, respectively. Equation 33

Table 8
Percentages of True, Universe, and Error Score Variances in Conventional and G-Theory Analyses

Design and Scale	Conventional						G-theory					
	True score	Specific factor	Transient	Random response	% Overestimate CE	% Overestimate CS	Universe score	Specific factor	Transient	Random response	% Overestimate CE	% Overestimate CS
$p \times I \times O$												
BFI Agreeableness	70.41	9.95	8.26	11.38	11.73	14.13	70.30	9.93	8.25	11.52	11.73	14.13
BFI Conscientiousness	73.04	10.17	7.38	9.41	10.11	13.93	73.02	10.17	7.38	9.43	10.11	13.93
BFI Extraversion	81.59	8.05	4.23	6.13	5.19	9.87	81.58	8.05	4.23	6.14	5.19	9.87
BFI Neuroticism	78.16	8.27	6.13	7.44	7.84	10.58	78.16	8.27	6.13	7.44	7.84	10.58
BFI Openness	76.34	11.29	5.09	7.27	6.67	14.79	75.61	11.19	5.04	8.17	6.67	14.79
BIDR IM	77.05	10.55	4.52	7.88	5.87	13.70	76.64	10.50	4.49	8.37	5.86	13.70
BIDR SDE	66.06	14.68	7.32	11.94	11.08	22.23	66.03	14.68	7.31	11.98	11.08	22.23
TSBI Form A	76.64	7.38	9.34	6.64	12.19	9.63	76.25	7.34	9.29	7.12	12.19	9.63
TSBI Form B	76.81	6.30	11.18	5.71	14.56	8.20	76.40	6.27	11.12	6.21	14.56	8.20
$p \times S \times O$												
BFI Agreeableness	73.79	6.57	9.04	10.60	12.24	8.90	73.67	6.56	9.02	10.75	12.24	8.90
BFI Conscientiousness	76.64	6.57	8.44	8.35	11.02	8.57	76.62	6.57	8.44	8.37	11.02	8.57
BFI Extraversion	85.15	4.49	3.39	6.97	3.98	5.28	85.14	4.49	3.38	6.98	3.98	5.28
BFI Neuroticism	81.48	4.95	6.78	6.79	8.32	6.08	81.48	4.95	6.78	6.79	8.32	6.08
BFI Openness	79.39	8.24	6.00	6.36	7.56	10.38	78.63	8.16	5.94	7.27	7.56	10.38
BIDR IM	80.85	6.75	5.67	6.73	7.02	8.35	80.42	6.71	5.64	7.22	7.02	8.35
BIDR SDE	69.72	11.03	8.30	10.95	11.91	15.81	69.68	11.02	8.30	11.00	11.91	15.81
TSBI Form A	79.20	4.82	10.89	5.09	13.75	6.08	78.80	4.79	10.84	5.57	13.75	6.08
TSBI Form B	80.04	3.06	12.83	4.07	16.03	3.82	79.62	3.04	12.76	4.57	16.03	3.82
$p \times S \times O$ (adj.)												
BFI Agreeableness	74.72	5.65	9.15	10.49	12.24	7.56	74.60	5.64	9.13	10.63	12.24	7.56
BFI Conscientiousness	77.60	5.61	8.55	8.24	11.02	7.23	77.58	5.61	8.55	8.27	11.02	7.23
$p \times F \times O$												
TSBI Form A/B	81.15	2.47	7.95	8.43	9.79	3.04	80.65	2.46	7.90	8.99	9.79	3.04

Note. % Overestimate = percentage in overestimating the Coefficient of Equivalence and Stability. CE = Coefficient of Equivalence; CS = Coefficient of Stability; BFI = Big Five Inventory; BIDR = Balanced Inventory of Desirable Responding; IM = Impression Management; SDE = Self-Deceptive Enhancement; TSBI = Texas Social Behavior Inventory; adj. = adjusted.

illustrates computation of a cut-score specific D-coefficient from the $p \times I \times O$ design for an IM score two standard deviations above the mean that might be used to flag for *faking good*. Note that this cut-score specific D-coefficient (.939) is noticeably higher than its corresponding global D-coefficient (.715) derived in Equation 31.

$$\text{Estimated D-coefficient}_{IM \text{ score two SDs} > \bar{Y}} = \frac{.588 + [(4.0482 - 5.8021)^2 - .0582]}{.588 + [(4.0482 - 5.8021)^2 - .0582] + \left(\frac{1.610}{20} + \frac{.035}{1} + \frac{1.284}{20 \times 1} + \frac{.964}{20} + \frac{.006}{1} + \frac{.006}{20 \times 1}\right)} = .939 \quad (33)$$

In Figure 1, we depict estimated D-coefficients for all possible cut-points along the complete 20- to 140-point range of the IM scale for the $p \times I \times O$ ($n'_i = 20$ and $n'_o = 1$) and $p \times S \times O$ ($n'_s = 2$ and $n'_o = 1$) designs. The horizontal scales in the figure have been transformed from the task to the total score metric (i.e., item scale × 20 and split scale × 2). As would be expected based on greater variability among items than between splits, splits-based designs yield higher dependability coefficients than items-based designs.

Cut-score specific D-coefficients were developed originally based on a conventional index created by Livingston (1972). Livingston's estimation formula, shown in Equation 34, was intended for single-occasion data, and therefore is applicable only to single-facet designs:

$$\text{Estimated Livingston coefficient} = \frac{\text{CE} \times \hat{\sigma}^2_{\bar{Y}} + (\bar{Y} - C)^2}{\hat{\sigma}^2_{\bar{Y}} + (\bar{Y} - C)^2}, \quad (34)$$

where CE is a one-facet conventional coefficient of equivalence (see conventional formulas for the $p \times I$, $p \times S$, $p \times S$ (adj.), or $p \times F$ designs in Table A1), $\bar{Y}$ is the mean of observed scores, C is the cut-score, and $\hat{\sigma}^2_{\bar{Y}}$ is the estimated variance of scores.

To address this limitation, we derived Equation 35 to extend Livingston's original coefficient to two-facet designs in which data are collected on two occasions:

$$\text{Two-Facet Estimated Livingston coefficient} = \frac{\text{CES} \times \sqrt{\hat{\sigma}^2_{O1} \times \hat{\sigma}^2_{O2}} + (\bar{Y} - C)^2}{\sqrt{\hat{\sigma}^2_{O1} \times \hat{\sigma}^2_{O2}} + (\bar{Y} - C)^2}, \quad (35)$$

where CES is a two-facet conventional coefficient of equivalence and stability (see conventional formulas for $p \times I \times O$, $p \times S \times O$, $p \times S \times O$ (adj.), or $p \times F \times O$ designs in Table A2), $\bar{Y}$ is the grand mean of observed scores, C is the cut-score, $\hat{\sigma}^2_{O1}$ is the estimated variance of scores on Occasion 1, and $\hat{\sigma}^2_{O2}$ is the estimated variance of scores on Occasion 2.

The primary difference between Livingston and D-coefficients is that Livingston coefficients do not take into account task or occasion effects on the absolute level of scores, and, as a result, are generally greater than corresponding D-coefficients. These relationships are apparent in the curves being higher for Livingston

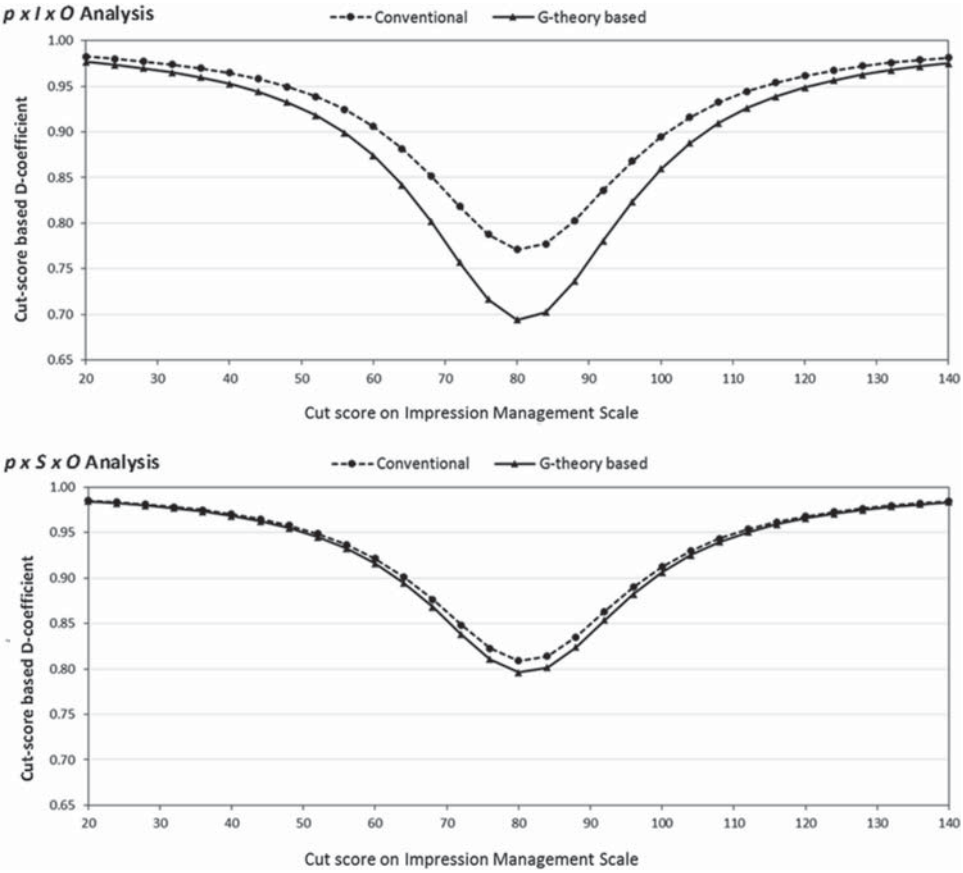

Figure 1. Cut-score based D-coefficients for the BIDR Impression Management Scale. Cut-scores are on the total score metric. Coefficients are given for both the $p \times I \times O$ (top) and $p \times S \times O$ (bottom) designs. G-Theory coefficient estimates are plotted with solid lines, and conventional Livingston-based coefficient estimates with dashed lines.

coefficients than for D-coefficients in Figure 1. The difference between Livingston and D-coefficients is larger for items than for splits because variability in means is greater among items than between splits.

Optimizing G- and D-Coefficients

After deriving appropriate indices of score consistency for a given situation, decision makers may find the coefficients unacceptably low or in need of improvement. Common ways to enhance reliability for designs illustrated here are to expand sampling of behaviors to include additional tasks or occasions. CTT provides a way to estimate such changes in reliability using the Spearman-Brown prophecy formula (Brown, 1910; Spearman, 1910) shown in Equation 36:

$$\text{Estimated Reliability} = \frac{n \times \text{Current Reliability}}{1 + (n - 1) \times \text{Current Reliability}}. \quad (36)$$

The value for n in Equation 36 can represent either tasks or occasions. The most common application of the formula is to derive a Spearman-Brown split-half reliability coefficient for a full-length measure based on two embedded parallel splits (i.e., $n = 2$), as shown in Equation 37:

$$\text{Spearman-Brown Split-half Reliability Estimate} = \frac{2 \times r_{Split\ 1,\ Split\ 2}}{1 + r_{Split\ 1,\ Split\ 2}}. \quad (37)$$

However, Equation 36 is more broadly applicable. For example, setting $n = 3$ could represent tripling the length of a measure or pooling results across three administrations, as illustrated in Equations 38 and 39 using the parallel-form and test–retest reliability coefficients for Form A of the TSBI reported in Table 4:

$$\text{Estimated Reliability for TSBI Form A tripled in length} = \frac{3 \times r_{Form\ A,\ Form\ B}}{1 + (3 - 1)\, r_{Form\ A,\ Form\ B}} = \frac{3 \times .8910}{1 + 2 \times .8910} = .961; \quad (38)$$

$$\text{Estimated Reliability for TSBI Form A pooled over three occasions} = \frac{3 \times r_{Occasion\ 1,\ Occasion\ 2}}{1 + (3 - 1)\, r_{Occasion\ 1,\ Occasion\ 2}} = \frac{3 \times .8402}{1 + 2 \times .8402} = .940. \quad (39)$$

Use of the Spearman-Brown prophecy formula assumes *classically parallel* measures and can be applied to either adding

tasks or occasions but not both at the same time. In contrast, G-theory techniques need not assume parallel measures and can gauge effects of changes to any facets of a design, separate or combined.

In the language of G-theory, defining and estimating the variance components for the universe of admissible observations is called a Generalizability, or G, study, and use of those components to estimate or optimize score consistency within a desired universe of generalization is called a Decision, or D, study. In D-studies, number of replicates can be specified for any given facet, and generalizability and dependability indices can be estimated for various combinations of fixed and random facets. Additionally, random facets from G-studies can be treated as fixed in D-studies (see, e.g., Equations 19 and 20), and non-nested facets can be treated as nested (see, e.g., Brennan, 2001a, pp. 15–17). For example, with the present data, D-study analyses to gauge effects of changing numbers of tasks and/or occasions would entail plugging those numbers into the formulas for G- and D-coefficients shown in the appendices for generalizing over tasks, occasions, or both.

To illustrate, consider again the results for Form A of the TSBI. In Table 7, the G-theory CES formula for the form-based analysis yielded a value of .807 using $n'_f = 1$ and $n'_o = 1$. Equation 40 illustrates estimation of reliability if the length of the TSBI is doubled and results pooled across two occasions (i.e., $n'_f = 2$ and $n'_o = 2$) while taking specific-factor, transient, and random-response error all into account:

$$\begin{aligned}\text{CES for TSBI}_{p\times F\times O \text{ with } n'_f=2 \text{ and } n'_o=2} &= \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \frac{\hat{\sigma}^2_{pf}}{n'_f} + \frac{\hat{\sigma}^2_{po}}{n'_o} + \frac{\hat{\sigma}^2_{pfo,e}}{n'_f n'_o}} \\ &= \frac{77.868}{77.868 + \frac{2.370}{2} + \frac{7.627}{2} + \frac{8.682}{2\times 2}} = .916. \end{aligned} \quad (40)$$

Note that doubling items and occasions raises the reliability estimate from .807 to .916.

Global and cut-score specific D-coefficients can be adjusted in a manner similar to G-coefficients using the formulas shown in Table A3 and Appendix B. To pinpoint how to optimize a measure under practical constraints, results for G- and global D-coefficients are commonly expanded and depicted in graphs like those shown in Figure 2 to simultaneously represent score consistency for a wide range of possible changes to a measurement procedure (see, e.g., Mushquash & O'Connor, 2006; Vispoel & Tao, 2013). The figure shows G- and global D-coefficients for the BFI Openness scale with numbers of items ranging from 10 to 30 and occasions from 1 to 3. Note that if scores are being used for norm referencing and a G-coefficient of at least .80 is desired, results for the original 10-item scale could be pooled across two occasions, or five comparable items could be added to the original scale on one occasion. As would typically be the case, global D-coefficients in Figure 2 are lower in magnitude than corresponding G-coefficients, but show a common trend. Both coefficients increase as numbers of items and/or occasions increase, but at a diminishing rate, as more and more items or occasions are added.

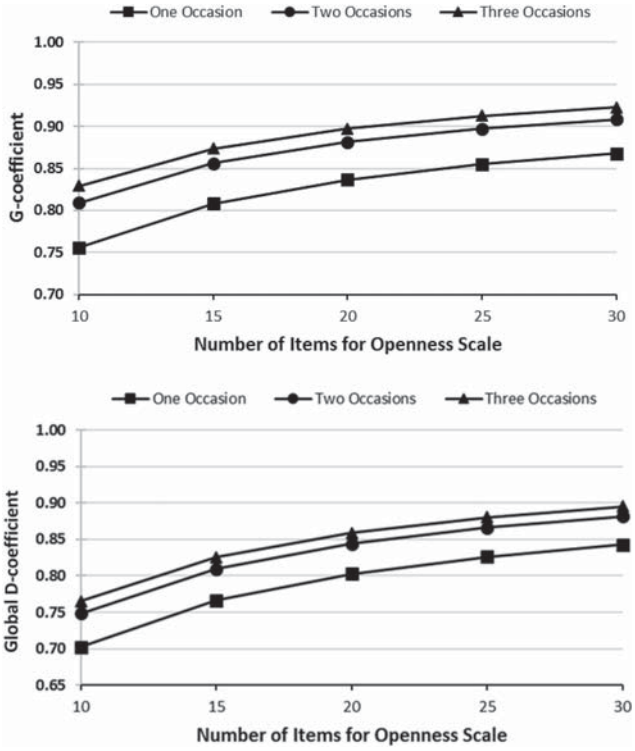

Figure 2. G-coefficient (top) and Global D-coefficient (bottom). Estimates for varying numbers of items and occasions for the BFI Openness Scale.

Disattenuating Validity Coefficients for Measurement Error

A serious problem with any reliability coefficient that does not account for the primary sources of measurement error affecting scores is that it can lead to misrepresentations of relationships between constructs when correcting correlation coefficients for measurement error. This problem is illustrated in Equation 41 using the classic *correction for attenuation* formula introduced by Spearman (1904):

$$\hat{\rho}_{T_X T_Y} = \frac{r_{XY}}{\sqrt{\hat{\rho}_{XX'} \times \hat{\rho}_{YY'}}}, \quad (41)$$

where $\hat{\rho}_{T_X T_Y}$ = estimated correlation between true scores for measures X and Y, r_{XY} = correlation between observed scores for measures X and Y, $\hat{\rho}_{XX'}$ = estimated reliability coefficient for measure X, and $\hat{\rho}_{YY'}$ = estimated reliability coefficient for measure Y. Note that inaccurate estimates of reliability in the denominator of the formula will result in either over- or underestimation of the correlation between true scores, thereby highlighting the need to include the most accurate available estimates of reliability when using it (i.e., CESs in the present context).

Although frequently overlooked, G-theory can provide an effective means to disattenuate correlations for measurement error when scores of interest are analyzed together in a multivariate design. Such disattenuated correlations can be derived for any G-theory design using Equation 42.

$$\hat{\rho}_{\nu\nu'}(p), _{G\text{-}theory} = \frac{\hat{\sigma}_{\nu\nu'}(p)}{\sqrt{\hat{\sigma}^2_{\nu}(p) \times \hat{\sigma}^2_{\nu'}(p)}}, \quad (42)$$

where $\hat{\rho}_{\nu\nu'}(p)$ represents the disattenuated correlation between scales ν and ν', $\hat{\sigma}_{\nu\nu'}(p)$ represents the estimated covariance of universe scores on scales ν and ν', and $\hat{\sigma}^2_{\nu}(p)$ and $\hat{\sigma}^2_{\nu'}(p)$ represent the estimated universe score variances on scales ν and ν', respectively.

Computing $\hat{\sigma}_{\nu\nu'}(p)$, $\hat{\sigma}^2_{\nu}(p)$, and $\hat{\sigma}^2_{\nu'}(p)$ can be complex (see, e.g., Brennan, 2001a), but these, along with disattenuated correlations, are routinely provided as output from mGENOVA (Brennan, 2001b). However, for the present two-facet designs, equivalent disattenuated correlations can be easily computed from G-theory and conventional CESs using Equations 43 and 44, respectively.

$$\hat{\rho}_{\nu\nu'}(p), _{G\text{-}theory} = \frac{[\widehat{\text{Cov}}(\text{Scale}X_{O1}, \text{Scale}Y_{O2}) + \widehat{\text{Cov}}(\text{Scale}X_{O2}, \text{Scale}Y_{O1})]/2}{\sqrt{\left(\text{CES}_{\text{Scale}X} \times \left(\hat{\sigma}^2_{\text{Scale}X_{O1}} + \hat{\sigma}^2_{\text{Scale}X_{O2}}\right)/2\right) \times \left(\text{CES}_{\text{Scale}Y} \times \left(\hat{\sigma}^2_{\text{Scale}Y_{O1}} + \hat{\sigma}^2_{\text{Scale}Y_{O2}}\right)/2\right)}} \quad (43)$$

$$\hat{\rho}_{T_X T_Y}, _{CTT} = \frac{[\widehat{\text{Cov}}(\text{Scale}X_{\text{O1}}, \text{Scale}Y_{\text{O2}}) + \widehat{\text{Cov}}(\text{Scale}X_{\text{O2}}, \text{Scale}\text{Y}_{\text{O1}})]/2}{\sqrt{\left(\text{CES}_{\text{Scale}X} \times \sqrt{\hat{\sigma}^2_{\text{Scale}X_{O1}} \times \hat{\sigma}^2_{\text{Scale}X_{O2}}}\right) \times \left(\text{CES}_{\text{Scale}Y} \times \sqrt{\hat{\sigma}^2_{\text{Scale}Y_{O1}} \times \hat{\sigma}^2_{\text{Scale}Y_{O2}}}\right)}}. \quad (44)$$

Because geometric and arithmetic means in the denominators of Equations 43 and 44 cancel out during computations, both formulas will yield the same result.

We illustrate similarities and differences between observed and disattenuated correlations based on hypothesized relationships between SDE and IM scores from the BIDR and selected subscale scores from the BFI and TSBI. In the research literature, SDE has been linked more strongly to Neuroticism and Self-Esteem/Social Competence, and IM more strongly to Agreeableness (see, e.g., Huang, 2013; Paulhus, 1991, 1999; Paulhus & Reid, 1991; Stöber, Dette, & Musch, 2002; Vispoel & Kim, 2014; Vispoel, Morris, & Kilinc, 2016). In Table 9, we provide an average across occasions of the original correlations between BIDR and BFI/TSBI scores using a method adapted from Le, Schmidt, and Putka (2009) to remove the effects of correlated transient error within occasions (see Footnote a in Table 9 for the formula). The disattenuated correlations for split- and item-based G-theory designs can be computed using either mGENOVA or Equations 43 and 44.

Computation of the disattenuated correlation between BIDR SDE(X) and BFI Neuroticism(Y) scores using Equation 43 for items is illustrated in Equation 45. The estimated covariances in the numerator were computed by multiplying the appropriate interscale correlation coefficients by the product of their corresponding standard deviations [e.g., $\widehat{\text{Cov}}(X_{O1}, Y_{O2}) = r_{X_{O1}, Y_{O2}} \times \hat{\sigma}_{X_{O1}} \times \hat{\sigma}_{Y_{O2}}$]. All values in Equation 45 come from Tables 3 and 7, except for the interscale correlations, which equal −.481 for $r_{X_{O1}, Y_{O2}}$ and −.405 for $r_{X_{O2}, Y_{O1}}$.

$$\hat{\rho}_{BIDR\ SDE,BFI\ Neuroticism}(p), _{G\text{-}theory} = \frac{[(-.481 \times 13.56 \times 6.35) + (-.405 \times 13.99 \times 6.30)]/2}{\sqrt{(.660 \times (13.56^2 + 13.99^2)/2) \times (.782 \times (6.30^2 + 6.35^2)/2)}} = -.616. \quad (45)$$

Table 9
Observed and Disattenuated Correlation Coefficients for Selected Scales

Method and Scale	Conventional		G-theory	
	BIDR IM	BIDR SDE	BIDR IM	BIDR SDE
Averaged two-occasion observed[a]				
Agreeableness	**.439**	.242		
Neuroticism	−.227	**−.443**		
TSBI Form A	.099	**.439**		
TSBI Form B	.002	**.378**		
Items-based disattenuated				
Agreeableness	**.596**	.354	**.596**	.354
Neuroticism	−.291	**−.616**	−.291	**−.616**
TSBI Form A	.128	**.618**	.128	**.618**
TSBI Form B	.003	**.531**	.003	**.531**
Splits-based disattenuated				
Agreeableness[b]	**.564**	.335	**.564**	.335
Neuroticism	−.279	**−.587**	−.279	**−.587**
TSBI Form A	.123	**.591**	.123	**.591**
TSBI Form B	.003	**.507**	.003	**.507**

Note. Convergent validity coefficients are shown in bold. BIDR = Balanced Inventory of Desirable Responding; IM = Impression Management; SDE = Self-Deceptive Enhancement; TSBI = Texas Social Behavior Inventory.
[a] Results were averaged across occasions using the following formula adapted from Le, Schmidt, and Putka (2009) to eliminate possible correlated transient error effects within occasions: $\frac{[\widehat{\text{Cov}}(\text{Scale}X_{O1}, \text{Scale}Y_{O2}) + \widehat{\text{Cov}}(\text{Scale}X_{O2}, \text{Scale}Y_{O1})]/2}{(\hat{\sigma}^2_{\text{Scale}X_{O1}} \times \hat{\sigma}^2_{\text{Scale}X_{O2}} \times \hat{\sigma}^2_{\text{Scale}Y_{O1}} \times \hat{\sigma}^2_{\text{Scale}Y_{O2}})^{1/4}}$, where $\widehat{\text{Cov}}$ is estimated covariance, $O1$ is occasion 1 and $O2$ is occasion 2. [b] Adjusted coefficients of equivalence and stability for uneven splits were used in equations 43 and 44 to derive disattenuated correlations for Agreeableness.

The complete results for convergent and discriminant validity shown in Table 9 reveal that median disattenuated convergent validity coefficients for items and splits are 0.17 and 0.14 higher, respectively, in both absolute and squared values than are corresponding observed-score correlations, thereby implying noticeably stronger relationships among the underlying constructs. Disattenuated correlations for items are higher than those for splits because their corresponding CESs are lower.

Additional Considerations When Using G-Theory

Estimating Variance Components

Computations of all G-theory indices discussed thus far were based on variance components estimated using mean squares from conventional ANOVA models. A problem with such estimates is that they will sometimes yield negative results, as shown in Tables 5 and 6 for split and occasion effects with the BIDR and TSBI scales. To address this problem, negative variance components can be set to zero at some point in the estimation process (see, e.g., Brennan, 2001a; Cronbach et al., 1972), or other methods can be used that do not yield negative components based on maximum likelihood estimation (Harville, 1977; Marcoulides, 1990; Searle, 1971) or Bayesian procedures (LoPilato, Carter, & Wang, 2015). Maximum likelihood (ML) and restricted maximum likelihood estimation (REML) options are readily available in popular software packages such as SPPS, SAS, and R.

In Table 10, we show results for variance components estimated using ANOVA mean squares, ML, and REML along with corresponding G- and global D-coefficients for BIDR and TSBI subscale scores from the $p \times S \times O$ design originally shown in Table 6, in which negative variance components were found for splits or occasions. We estimated these variance components using the VARCOMP procedure in SAS 9.3. Note that the ML and REML procedures eliminated all negative variance components, but had little effect on G- and global D-coefficients. In other situations, particularly with larger negative variance component estimates, smaller sample sizes, unbalanced designs, or missing data, differences may be more noteworthy (Raykov & Marcoulides, 2011).

Assumptions of G-Theory

To properly interpret reliability coefficients from G-theory designs, careful attention should be given to underlying assumptions. We consider four important assumptions related to sampling, independence, measurement scales, and structure of scores.

Sampling. A key assumption stated at the outset of this article was that tasks and occasions were sampled at random from associated universes of admissible observations. Because this will rarely be literally true in practice, the notion of *exchangeability* is routinely invoked as a reasonable substitute for strict random sampling (de Finetti, 1937). From this perspective, we could consider a facet (e.g., tasks, occasions) as random if conditions not observed in the G-study are exchangeable with those that were observed. Assuming exchangeability is reasonable in the present context because alternative tasks and occasions could serve the same purposes.

Independence. The independence assumption does not mean that scores are uncorrelated with each other. Rather, it means that the experience of answering part of an assessment (e.g., a given item) does not affect responses to any other part of the assessment (e.g., other items), and that a given individual's scores have no bearing on another individual's scores. An alternative way of stating this assumption is that error scores for parts of an assess-

Table 10
Alternative Variance Components, G-Coefficients, and Global D-Coefficients for Selected Scales

	Estimation method				Estimation method		
Scale, Effects, or Index	ANOVA mean squares	ML	REML	Scale, Effects, or Index	ANOVA mean squares	ML	REML
IM splits				SDE splits			
person (*p*)	61.728	61.541	61.758	person (*p*)	33.066	33.002	33.095
split (*s*)	−.060	.000	.000	split (*s*)	−.058	.000	.000
occasion (*o*)	.642	.427	.628	occasion (*o*)	.899	.532	.883
$p \times s$	10.305	10.245	10.245	$p \times s$	10.458	10.400	10.400
$p \times o$	4.331	4.347	4.345	$p \times o$	3.938	3.957	3.955
$s \times o$	−.029	.000	.000	$s \times o$	−.033	.000	.000
$p \times s \times o$, *error*	11.084	11.055	11.055	$p \times s \times o$, *error*	10.435	10.402	10.402
G-coefficient	.804	.804	.805	G-coefficient	.697	.697	.697
Global D-coefficient	.798	.800	.798	Global D-coefficient	.684	.689	.685
TSBI A Splits				TSBI B Splits			
person (*p*)	18.161	18.071	18.174	person (*p*)	20.086	19.984	20.086
split (*s*)	−.011	.000	.000	split (*s*)	.032	.022	.027
occasion (*o*)	−.012	.000	.000	occasion (*o*)	−.001	.000	.000
$p \times s$	2.210	2.199	2.199	$p \times s$	1.536	1.542	1.541
$p \times o$	2.498	2.485	2.485	$p \times o$	3.220	3.218	3.218
$s \times o$	−.006	.000	.000	$s \times o$	−.011	.000	.000
$p \times s \times o$, *error*	2.570	2.563	2.563	$p \times s \times o$, *error*	2.305	2.294	2.294
G-coefficient	.788	.788	.789	G-coefficient	.796	.796	.796
Global D-coefficient	.788	.788	.789	Global D-coefficient	.796	.795	.796

Note. ML = maximum likelihood estimation; REML = restricted maximum likelihood estimation; IM = Impression Management; SDE = Self-Deceptive Enhancement; TSBI = Texas Social Behavior Inventory.

ment are uncorrelated with error scores from other parts and are uncorrelated across persons. This assumption is shared by both G-theory and CTT.

Measurement scale. Another assumption common to all G-theory and CTT applications illustrated thus far is that the underlying variable of interest is continuous and measured on an equal-interval scale. This assumption is implicit in the use of means and variances in deriving G-theory coefficients. The same assumption along with linear relationships between variables is made when interpreting correlation coefficients using G-theory or CTT.

Underlying structure of scores. When comparing G-theory and CTT, we noted that the G-coefficient for the $p \times I$ design is equivalent to coefficient alpha. As a result, this G-coefficient will be susceptible to the same criticisms levied on alpha in recent years (see, e.g., Becker, 2000; Dunn, Baguley, & Brunsden, 2014; Green & Yang, 2009; Schmidt et al., 2003; Schmitt, 1996; Sijtsma, 2009). The primary criticism is that alpha will be a suitable index of reliability for a single occasion only when items yield scores that are *essentially tau-equivalent*. With essential tau-equivalence, error and observed scores can vary freely among items, but true scores can differ only by an additive constant. When this assumption holds, error and observed-score variances can differ across items but true-score variances cannot. In practice, this assumption is unlikely to be met because of the diversity of items typically included in a measure.

By definition, G-theory coefficients (with alpha as a special case) represent reliability for *randomly parallel* measures, as indexed by the average consistency of scores across the universe of interest (Cronbach, Rajaratnam, & Gleser, 1963). Accordingly, such coefficients reflect conservative estimates of reliability unless scores are essentially tau-equivalent or classically parallel. We also noted earlier that Spearman-Brown split-half, parallel-form, and test-retest coefficients will equal their corresponding G-coefficients only when task or occasion scores have equal variances. If not, G-coefficients will be lower because arithmetic rather than geometric means are used in the denominators of computational formulas (see Table A1 and A2).

Structural Equation Modeling Alternatives to G-Theory

Over recent years, researchers have used confirmatory factor analysis (CFA) models to compute score reliability in both traditional and new ways. To represent essential tau-equivalence in a factor model, each item would have the same unstandardized factor loading on a single general factor. With such a model, the proportion of variation in scores accounted for by that single factor would equal coefficient alpha. However, CFA models need not be restricted to identical loadings nor limited to single factors. Omega coefficients represent a more recently developed general class of reliability estimates that reflect proportions of variation in scores accounted for by factor models (McDonald, 1999; Zinbarg, Revelle, Yovel, & Li, 2005).

The omega coefficient currently reported most often in the literature is one representing *congeneric* scores on a single general factor. Factor modeling for this coefficient is similar to that for alpha except that unstandardized factor loadings are allowed to vary. An advantage this coefficient shares with alpha is that it produces a unique value for a given set of data. In Table 11, we report estimated alpha and omega coefficients computed using the *psych* package (Revelle, 2015) from R Version 3.2.3. Note that omegas exceed alphas for all subscales. Within the population, omegas will be greater than or equal to alphas and coincide only when scores for items are essentially tau-equivalent (Zinbarg et al., 2005).

CFA models for essentially tau-equivalent measures also can be used to derive variance components (Marcoulides, 1990, 1996; Raykov & Marcoulides, 2006). Although conceptually enlightening, such models would offer little practical advantage over methods described previously because they yield the same results for G-coefficients and do not produce variance components needed to derive D-coefficients as readily. However, alternative CFA designs based on latent state-trait theory (Steyer, Ferring, & Schmitt, 1992; Steyer, Mayer, Geiser, & Cole, 2015) and multitrait-multimethod methodology (Eid et al., 2008) allow for less restricted estimation of variance for persons, states, traits, methods, occasions, and residuals on the level of items, splits, and scales.

Table 11
Alternative Reliability Coefficients for $p \times I$ and $p \times I \times O$ Designs

	$p \times I$				$p \times I \times O$	
Scale	Alpha	Ordinal alpha	Omega	Ordinal omega	Equal-interval G-coefficient	Ordinal G-coefficient
BFI						
Agreeableness	.756	.811	.769	.814	.703	.741
Conscientiousness	.793	.836	.810	.838	.730	.754
Extraversion	.853	.881	.859	.887	.816	.825
Neuroticism	.837	.864	.839	.866	.782	.803
Openness	.774	.806	.800	.817	.756	.788
BIDR						
IM	.781	.820	.794	.826	.766	.783
SDE	.711	.734	.723	.739	.660	.668
TSBI						
Form A	.837	.860	.841	.864	.763	.780
Form B	.864	.879	.873	.885	.764	.772

Note. BFI = Big Five Inventory; BIDR = Balanced Inventory of Desirable Responding; IM = Impression Management; SDE = Self-Deceptive Enhancement; TSBI = Texas Social Behavior Inventory.

These comprehensive procedures permit integration of congeneric measures in testing complex models combining more intricate analysis of reliability and construct validity (see, e.g., Geiser et al., 2015). We look forward to expanded use of these emerging techniques in coming years.

Ordinal G-Theory

G-Theory analyses described until now have treated continuous variables as being measured on equal-interval scales. However, strictly speaking, the Likert-style scales within the present questionnaires yield scores for ordered but not necessarily equally spaced categories on the response metric. Recent developments in software and measurement models allow for extensions of G-theory to ordinal scales in deriving variance components and corresponding reliability estimates (see, e.g., Ark, 2015; Gadermann, Guhn, & Zumbo, 2012; Zumbo, Gadermann, & Zeisser, 2007). The procedures are similar to those for the basic CFA models discussed already except that polychoric rather than conventional covariance or correlation matrices for items are analyzed, and results are interpreted in relation to latent variables rather than observed scores.

In Table 11, we provide ordinal alpha and ordinal omega coefficients for the $p \times I$ design and G-coefficients (CESs) for the ordinal and equal-interval $p \times I \times O$ designs. We computed ordinal alpha and ordinal omega coefficients from polychoric correlation matrices produced from the *psych* package (Revelle, 2015) in R and obtained G-coefficients for the ordinal $p \times I \times O$ design by inputting the polychoric correlation matrices from R into Mplus Version 7.4 using the factor model for a two-facet, crossed design described by Raykov and Marcoulides (2006). For our results in Table 11, ordinal omegas always exceed ordinal alphas, and both coefficients exceed their equal-interval counterparts. Similar patterns are evident with the two-facet designs with G-coefficients for the ordinal $p \times I \times O$ design always exceeding those for the equal-interval $p \times I \times O$ design. Differences between ordinal and equal-interval reliability coefficients result in part from ordinal indices being referenced to a latent variable rather than observed score metric. Such differences are generally more pronounced for scales having fewer response alternatives and diminish as more options are added (Ark, 2015; Rhemtulla, Brosseau-Liard, & Savalei, 2012).

Software for Doing G-Theory Analyses

Readers interested in applying G-theory techniques illustrated here can take advantage of several statistical packages listed in Table 12, most of which can be obtained free of charge from websites identified in the table. We also indicate whether each package provides sufficient information to conduct the analyses demonstrated here. All packages allow for input of raw data and produce output for balanced, multifaceted designs. The programs differ in alternative data entry options (variance components, mean squares); nature of the user interface (menu vs. command driven); ability to handle D-studies, unbalanced designs, missing data, and large data sets; and options for extended output (confidence intervals for variance components, global and cut-score specific D-coefficients, and disattenuated validity coefficients).

Additional resources are available for readers to derive variance components, compute omega coefficients, and conduct G-theory analyses for ordinal-level data. Variance components for G-theory analyses can be computed using the VARCOMP procedure in SPSS and PROC VARCOMP or PROC MIXED procedures in SAS (also see Mushquash & O'Connor, 2006); omega coefficients can be derived using the *psych* package (Revelle, 2015) in R; and structural equation modeling programs such as Mplus, AMOS, EQS, and LISREL can provide variance components and G-coefficients for $p \times I$ and $p \times I \times O$ designs for both equal-interval and ordinal scales.

Table 12
Features of Computer Packages for Performing G-Theory Analyses

Characteristic	Option	EduG	G_string IV[a]	GENOVA	mGENOVA	urGENOVA
Data input	Score-level data with no missing values	✓	✓	✓	✓	✓
	Score-level data with missing values		✓			✓
	Variance components			✓		
	Mean squares	✓		✓	✓	
Interface	Menu based	✓	✓			
	Command line based			✓	✓	✓
Analysis	G-study	✓	✓	✓	✓	✓
	D-study	✓	✓	✓	✓	
Design	Balanced multifaceted	✓	✓	✓	✓	✓
	Unbalanced multifaceted				✓	✓
Analyses relevant to those presented in this article	Variance components	✓	✓	✓	✓	✓
	Confidence intervals for variance components		✓			✓
	G-coefficients	✓	✓	✓	✓	
	Global D-coefficients	✓	✓	✓	✓	
	Cut-score-based D-coefficients	✓		✓		
	Disattenuated correlations				✓	
	Composite reliability coefficients				✓	

Note. GENOVA, urGENOVA, and mGENOVA can be downloaded from http://www.education.uiowa.edu/centers/casma/computer-programs; G String IV from http://fhsperd.mcmaster.ca/g_string; and SPSS and SAS code for computing variance components from https://people.ok.ubc.ca/brioconn/gtheory/gtheory.html. EduG is available from Cardinet, Johnson, and Pini (2010).
[a] G string IV is a graphical user interface for urGENOVA that also provides extended output. It is limited to a maximum of 1,500 subjects.

Discussion

Our primary motivation in writing this article was to promote greater understanding and use of G-theory for evaluating psychometric properties of scores for measures of individual differences. We illustrated connections between G-theory and CTT in greater detail than has been done previously, and in so doing, demonstrated that CTT truly is a special case of G-theory when scores are classically parallel. However, G-theory is not restricted to classically parallel measures and offers further advantages over CTT in isolating multiple sources of measurement error, deriving dependability coefficients for cut-scores, and optimizing a measurement procedure.

Sources of Measurement Error

In demonstrating G-theory's effectiveness in estimating multiple sources of measurement error, we stressed the importance of integrating specific-factor, transient, and random-response error into indices of score consistency. Separating these estimates revealed how ignoring particular sources of measurement error can inflate score consistency, pinpointed the type of error most prevalent in a given situation, and allowed for estimation of score consistency when altering conditions for facets in a design. In CTT, the most common way to take specific-factor, transient, and random-response error into account is by deriving a CES based on administration of different parallel forms on different occasions. However, with this approach, two forms are needed, and sources of measurement error are confounded within a single error term. To separate specific-factor, transient, and random-response error, the same measure or measures need to be administered on at least two occasions.

We considered three G-theory designs that satisfy these conditions: $p \times F \times O$, $p \times S \times O$, and $p \times I \times O$. Although potentially most informative, the $p \times F \times O$ design shares the drawback of CTT-CESs in requiring two forms plus the added burden of administering both forms on two occasions. As a result, this design seems unlikely to be used much in practice. We examined the $p \times S \times O$ design primarily to show its relationship with CTT, but at the same time demonstrated that it might offer an attractive alternative to the $p \times F \times O$ design by combining modeling of parallel tasks with the practicality of the $p \times I \times O$ design in requiring only one form (see Vispoel et al., 2016, for further evidence of possible benefits when using split-measures in G-theory analyses).

Dependability of Scores for Criterion-Referenced Decisions

Another valuable and underused aspect of G-theory is in estimating D-coefficients targeted directly to cut-scores used in decision making. Although these coefficients have no direct counterparts in CTT, they bear strong similarities to Livingston's (1972) dependability coefficient. G-theory-based D-coefficients are preferable to Livingston's coefficient, because they extend beyond single-occasion data, take absolute error into account, and include a correction for bias.

We illustrated a very specialized application of D-coefficients in providing evidence of dependability for cut-scores on the BIDR IM scale that might be used to flag instances of possible faking. However, D-coefficients are relevant to a much wider variety of applications encompassing use of cut-scores in making screening, selection, admission, and classification decisions in employment, academic, and other settings. In these situations, D-coefficients could be reported at particular points or score ranges over which decisions are made.

Optimizing Score Consistency

Another key advantage of G-theory is in providing comprehensive estimates of how very specific changes in a measurement procedure might impact score consistency. Such estimates are also very easy to compute by specifying numbers of replications in appropriate formulas for any individual or combination of facets in a design. In the present context, this was accomplished simply by changing n' values for tasks and occasions in the equations for estimating G- and D-coefficients provided in the appendices.

However, before deciding how to alter a measurement procedure, careful attention should be given to how variance components are estimated. For example, when deriving variance components for completely crossed, G-theory designs with two random facets, Smith (1978) recommends that the product of n' values for persons and the measurement facets (e.g., $n'_p \times n'_i \times n'_o$) be no lower than 800. Also, with small sample sizes, dichotomous data, unbalanced designs, or missing data, maximum-likelihood-based procedures often provide more accurate variance components than do ANOVA mean squares (Raykov & Marcoulides, 2011). Unfortunately, in practice, such criteria are often not met or are ignored altogether.

When variance components are properly estimated, further insights into how to change a measure might be gained by examining relative proportions of measurement error. For example, within the $p \times T \times O$ design, large proportions of specific-factor error would signal a need to lengthen a measure, and large proportions of transient error to pool results across occasions. However, before making such changes, attention should be given to the way in which tasks are represented and the feasibility of making particular changes. Because of greater specific-factor error, traditional item-level G-theory analyses might suggest unnecessary increases in items, and practicality may not allow for repeated administrations of measures.

Software Concerns

A major stumbling block to using G-theory has been the lack of readily available software to conduct the analyses. This has also hindered reporting of alternative indices of reliability for single occasions such as omega and ordinal alpha coefficients. We hope that computer resources described here facilitate greater applications of G-theory and better reporting of reliability in general. In absence of these resources, G-theory indices for the present designs can be computed from formulas provided in the appendices, and when appropriate, alternative variance components can be substituted into the formulas to derive more accurate estimates of score consistency and disattenuated validity coefficients.

Conclusions

An important goal in writing this article was to provide a strong foundation for understanding G-theory by linking it to CTT in a more comprehensive manner than has been done in the past. Formulas presented and contrasted in the text and appendices should help readers in moving from one theory to the other and in

recognizing how, when, and why indices across theories align or differ. We also hope that concepts discussed here enhance understanding of more complicated G-theory designs, incorporation of other facets of measurement into those designs (item parcels, testlets, task orderings, modes of administration, judges, prompts, etc.), and connections between G-theory and CFA in modeling reliability and construct validity of scores. We encourage readers to review the many excellent writings cited throughout this article for additional applications of G-theory to take further advantage of its unrealized potential for assessing psychometric properties of scores, gauging interrelations among psychological constructs, and determining effective ways to improve measurement procedures.

References

Ark, T. K. (2015, June 11). *Ordinal generalizability theory using an underlying latent variable framework* (Unpublished doctoral dissertation). University of British Columbia, Vancouver, Canada. http://dx.doi.org/10.14288/1.0166304

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods, 5,* 370–379. http://dx.doi.org/10.1037/1082-989X.5.3.370

Brennan, R. L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement, 22,* 307–331. http://dx.doi.org/10.1177/014662169802200401

Brennan, R. L. (2001a). *Generalizability theory*. New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4757-3456-0

Brennan, R. L. (2001b). *Manual for mGENOVA (Version 2.1)*. Iowa City, IA: University of Iowa, IA Testing Programs. (Iowa Testing Programs Occasional Papers No. 50)

Brennan, R. L. (2001c). *Manual for urGENOVA (Version 2.1)*. Iowa City, IA: University of Iowa, IA Testing Programs.

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14,* 277–289. http://dx.doi.org/10.1111/j.1745-3984.1977.tb00045.x

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296–322. http://dx.doi.org/10.1111/j.2044-8295.1910.tb00207.x

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. New York, NY: Routledge.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334. http://dx.doi.org/10.1007/BF02310555

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

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

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16,* 137–163. http://dx.doi.org/10.1111/j.2044-8317.1963.tb00206.x

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64,* 391–418. http://dx.doi.org/10.1177/0013164404266386

de Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives [Forecast/prediction: Its logical laws, its subjective sources]. *Annales de l'institut Henri Poincaré, 7,* 1–68. Retrieved from https://eudml.org/doc/79004 (Published in English in *Studies in Subjective Probability,* edited by H. E. Kyburg Jr., & H. Smokler, New York, NY, Wiley, 1964, pp. 95–158).

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105,* 399–412. http://dx.doi.org/10.1111/bjop.12046

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods, 13,* 230–253. http://dx.doi.org/10.1037/a0013219

Feldt, L. S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement, 44,* 883–891. http://dx.doi.org/10.1177/0013164484444010

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. H. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York, NY: Macmillan.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation, 17,* 1–13.

Geiser, C., Keller, B. T., Lockhart, G., Eid, M., Cole, D. A., & Koch, T. (2015). Distinguishing state variability from trait change in longitudinal data: The role of measurement (non)invariance in latent state-trait analyses. *Behavior Research Methods, 47,* 172–203. http://dx.doi.org/10.3758/s13428-014-0457-z

Green, S. B. (2003). A coefficient alpha for test-retest data. *Psychological Methods, 8,* 88–101. http://dx.doi.org/10.1037/1082-989X.8.1.88

Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika, 74,* 121–135. http://dx.doi.org/10.1007/s11336-008-9098-4

Haertel, E. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: Praeger.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association, 72,* 320–338. http://dx.doi.org/10.1080/01621459.1977.10480998

Helmreich, R., & Stapp, J. (1974). Short forms of the Texas Social Behavior Inventory (TSBI), an objective measure of self-esteem. *Bulletin of the Psychonomic Society, 4,* 473–475. http://dx.doi.org/10.3758/BF03334260

Huang, C. (2013). Relation between self-esteem and socially desirable responding and the role of socially desirable responding in the relation between self-esteem and performance. *European Journal of Psychology of Education, 28,* 663–683. http://dx.doi.org/10.1007/s10212-012-0134-5

Jarjoura, D. (1986). An estimator of examinee-level measurement error variance that considers test form difficulty adjustments. *Applied Psychological Measurement, 10,* 175–186. http://dx.doi.org/10.1177/014662168601000209

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory—Versions 4a and 54*. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research. Retrieved from http://www.outofservice.com/bigfive

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,* 165–200. http://dx.doi.org/10.1177/1094428107302900

Livingston, S. A. (1972). Criterion-referenced applications of classical test theory. *Journal of Educational Measurement, 9,* 13–26. http://dx.doi.org/10.1111/j.1745-3984.1972.tb00756.x

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

Lord, F. M. (1957). A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. *Psychometrika, 22,* 207–220. http://dx.doi.org/10.1007/BF02289122

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika, 30,* 239–270. http://dx.doi.org/10.1007/BF02289490

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

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

Marcoulides, G. A. (1996). Estimating variance components in generalizability theory: The covariance structure analysis approach [Teacher's corner]. *Structural Equation Modeling, 3,* 290–299. http://dx.doi.org/10.1080/10705519609540045

Marcoulides, G. A. (2000). Generalizability theory. In H. Tinsley & S. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 527–551). San Diego, CA: Academic Press. http://dx.doi.org/10.1016/B978-012691360-6/50019-7

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Erlbaum.

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38,* 542–547. http://dx.doi.org/10.3758/BF03192810

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press. http://dx.doi.org/10.1016/B978-0-12-590241-0.50006-X

Paulhus, D. L. (1999). *Paulhus Deception Scales (PDS): The Balanced Inventory of Desirable Responding-7* (User's manual). Toronto, Ontario, Canada: Multi-Health Systems.

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology, 60,* 307–317. http://dx.doi.org/10.1037/0022-3514.60.2.307

Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika, 42,* 549–565. http://dx.doi.org/10.1007/BF02295978

Raykov, T., & Marcoulides, G. A. (2006). Estimation of generalizability coefficients via a structural equation modeling approach to scale reliability evaluation. *International Journal of Testing, 6,* 81–95. http://dx.doi.org/10.1207/s15327574ijt0601_5

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Routledge.

Revelle, W. (2015). psych: Procedures for personality and psychological research (1.5.8) [Computer software package and manual]. Evanston, IL: Northwestern University. Retrieved from https://cran.r-project.org/web/packages/psych

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17,* 354–373. http://dx.doi.org/10.1037/a0029315

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review, 9,* 99–103.

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

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350–353. http://dx.doi.org/10.1037/1040-3590.8.4.350

Searle, S. R. (1971). *Linear models*. New York, NY: Wiley.

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

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44,* 922–932. http://dx.doi.org/10.1037/0003-066X.44.6.922

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107–120. http://dx.doi.org/10.1007/s11336-008-9101-0

Smith, P. L. (1978). Sampling error of variance components in small sample multifacet generalizability designs. *Journal of Educational Measurement, 3,* 319–346.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101. (Reprinted in *International Journal of Epidemiology, 39,* pp. 1137–1150, 2010).http://dx.doi.org/10.1093/ije/dyq191

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295. http://dx.doi.org/10.1111/j.2044-8295.1910.tb00206.x

Steyer, R., Ferring, D., & Schmitt, M. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment, 8,* 79–98.

Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and traits—Revised. *Annual Review of Clinical Psychology, 11,* 71–98. http://dx.doi.org/10.1146/annurev-clinpsy-032813-153719

Stöber, J., Dette, D. E., & Musch, J. (2002). Comparing continuous and dichotomous scoring of the balanced inventory of desirable responding. *Journal of Personality Assessment, 78,* 370–389. http://dx.doi.org/10.1207/S15327752JPA7802_10

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560–620). Washington, DC: American Council on Education.

Vispoel, W. P., & Kim, H. Y. (2014). Psychometric properties for the Balanced Inventory of Desirable Responding: Dichotomous versus polytomous conventional and IRT scoring. *Psychological Assessment, 26,* 878–891. http://dx.doi.org/10.1037/a0036430

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2016). Using G-theory to enhance evidence of reliability and validity for common uses of the Paulhus Deception Scales. *Assessment*. Advance online publication. http://dx.doi.org/10.1177/1073191116641182

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

Wiley, E. W., Webb, N. M., & Shavelson, R. J. (2013). The generalizability of test scores. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology: Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 43–60). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/14047-003

Woodruff, D. (1991). Stepping up test score conditional variances. *Journal of Educational Measurement, 28,* 191–196. http://dx.doi.org/10.1111/j.1745-3984.1991.tb00353.x

Woodruff, D., Traynor, A., Cui, Z., & Fang, Y. (2013, October). *A comparison of three methods for computing scale score conditional standard errors of measurement* (ACT Research Report Series 7). Iowa City, IA: ACT.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70,* 123–133. http://dx.doi.org/10.1007/s11336-003-0974-7

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods: JMASM, 6,* 21–29. Retrieved from http://digitalcommons.wayne.edu/jmasm

(*Appendices follow*)

Appendix A

G-Theory Coefficients and Their Conventional Counterparts

Table A1
Formulas for G-Theory and Conventional Reliability Coefficients for Single-Facet Designs

G-coefficient formula[a]	Conventional coefficient	Conversion
$p \times I$		
$\text{G-coefficient}_{p\times I} = \dfrac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \dfrac{\hat{\sigma}_{pi,e}^2}{n_i'}}$	Coefficient Alpha $= \dfrac{\left(\dfrac{n_{Total}}{n_{Total}-1}\right)\sum_{i}^{n_{Total}}\sum_{i\neq j}^{n_{Total}}\widehat{\text{Cov}}(\textit{Item i, Item j})}{\hat{\sigma}_Y^2}$	$\text{G-coefficient}_{p\times I}$ = Coefficient Alpha
$p \times S$		
$\text{G-coefficient}_{p\times S} = \dfrac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \dfrac{\hat{\sigma}_{ps,e}^2}{n_s'}}$	Rulon Split-Half $= \dfrac{4 \times \widehat{\text{Cov}}(S1, S2)}{\hat{\sigma}_Y^2}$	$\text{G-coefficient}_{p\times S}$ = Rulon Split-Half
$p \times S$ (adjusted)		
$\text{G-coefficient}_{p\times S(\text{adjusted})} = .25 \times \left(\dfrac{n_{Total}}{n_{Split1}} \times \dfrac{n_{Total}}{n_{Split2}}\right) \times \dfrac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \dfrac{\hat{\sigma}_{ps,e}^2}{n_s'}}$	Raju Split-Half $= \dfrac{\left(\dfrac{n_{Total}}{n_{Split\,1}} \times \dfrac{n_{Total}}{n_{Split\,2}}\right) \times \widehat{\text{Cov}}(S1, S2)}{\hat{\sigma}_Y^2}$	$\text{G-coefficient}_{p\times S\,(\text{adjusted})}$ = Raju Split-Half
$p \times F$		
$\text{G-coefficient}_{p\times F} = \dfrac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \dfrac{\hat{\sigma}_{pf,e}^2}{n_f'}}$	Parallel Forms Correlation $= \dfrac{\widehat{\text{Cov}}(F1, F2)}{\sqrt{\hat{\sigma}_{F1}^2 \times \hat{\sigma}_{F2}^2}}$	$\text{G-coefficient}_{p\times F} = \dfrac{(\text{Forms correlation}) \times \sqrt{\hat{\sigma}_{F1}^2 \times \hat{\sigma}_{F2}^2}}{(\hat{\sigma}_{F1}^2 + \hat{\sigma}_{F2}^2)/2} = \dfrac{\widehat{\text{Cov}}(F1, F2)}{(\hat{\sigma}_{F1}^2 + \hat{\sigma}_{F2}^2)/2}$
$p \times O$		
$\text{G-coefficient}_{p\times O} = \dfrac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \dfrac{\hat{\sigma}_{po,e}^2}{n_o'}}$	Test-retest $= \dfrac{\widehat{\text{Cov}}(O1, O2)}{\sqrt{\hat{\sigma}_{O1}^2 \times \hat{\sigma}_{O2}^2}}$	$\text{G-coefficient}_{p\times O} = \dfrac{(\text{Test-retest}) \times \sqrt{\hat{\sigma}_{O1}^2 \times \hat{\sigma}_{O2}^2}}{(\hat{\sigma}_{O1}^2 + \hat{\sigma}_{O2}^2)/2} = \dfrac{\widehat{\text{Cov}}(O1, O2)}{(\hat{\sigma}_{O1}^2 + \hat{\sigma}_{O2}^2)/2}$

Note. n_{Total} = number of items in full scale; $n_{Split\ 1}$ = number of items in split 1; $n_{Split\ 2}$ = number of items in split 2; Y = observed scores for a given occasion; $S1$ = Split 1 scores; $S2$ = Split 2 scores; $O1$ = Occasion 1 scores; $O2$ = Occasion 2 scores; $F1$ = Form 1 scores; $F2$ = Form 2 scores.
[a] For the present examples, n_o' and n_f' both equal 1, n_s' equals 2, and n_i' values range from 8 to 20.

(*Appendices continue*)

Table A2
Formulas for G-Theory and Conventional Reliability Coefficients for Two-Facet Designs

G-coefficient formula[a]	Conventional coefficient[b]
$p \times I \times O \text{ CE} = \dfrac{\hat{\sigma}^2_p + \dfrac{\hat{\sigma}^2_{po}}{n'_o}}{\hat{\sigma}^2_p + \dfrac{\hat{\sigma}^2_{pi}}{n'_i} + \dfrac{\hat{\sigma}^2_{po}}{n'_o} + \dfrac{\hat{\sigma}^2_{pio,e}}{n'_i n'_o}}$	Items-based Two Facet CE $= \dfrac{\left(\dfrac{n_{Total}}{2 \times (n_{Total}\text{-}1)}\right) \times \left[\left(\sum_i^{n_{Total}} \sum_{i \neq j}^{n_{Total}} \widehat{\text{Cov}}(I_i O_1, I_j O_1)\right) + \left(\sum_i^{n_{Total}} \sum_{i \neq j}^{n_{Total}} \widehat{\text{Cov}}(I_i O_2, I_j O_2)\right)\right]}{\sqrt{\hat{\sigma}^2_{O1} \times \hat{\sigma}^2_{O2}}}$
$p \times S \times O \text{ CE} = \dfrac{\hat{\sigma}^2_p + \dfrac{\hat{\sigma}^2_{po}}{n'_o}}{\hat{\sigma}^2_p + \dfrac{\hat{\sigma}^2_{ps}}{n'_s} + \dfrac{\hat{\sigma}^2_{po}}{n'_o} + \dfrac{\hat{\sigma}^2_{pso,e}}{n'_s n'_o}}$	Splits-based Two Facet CE $= \dfrac{2 \times (\widehat{\text{Cov}}(S_1 O_1, S_2 O_1) + \widehat{\text{Cov}}(S_1 O_2, S_2 O_2))}{\sqrt{\hat{\sigma}^2_{O1} \times \hat{\sigma}^2_{O2}}}$
$p \times S \times O$ CE (adjusted) $= .25 \times \left(\dfrac{n_{Total}}{n_{Split\,1}} \times \dfrac{n_{Total}}{n_{Split\,2}}\right) \times \left(\dfrac{\hat{\sigma}^2_p + \dfrac{\hat{\sigma}^2_{po}}{n'_o}}{\hat{\sigma}^2_p + \dfrac{\hat{\sigma}^2_{ps}}{n'_s} + \dfrac{\hat{\sigma}^2_{po}}{n'_o} + \dfrac{\hat{\sigma}^2_{pso,e}}{n'_s n'_o}}\right)$	Splits-based Two Facet CE (adjusted) $= \dfrac{.5 \times \left(\dfrac{n_{Total}}{n_{Split\,1}} \times \dfrac{n_{Total}}{n_{Split\,2}}\right) \times (\widehat{\text{Cov}}(S_1 O_1, S_2 O_1) + \widehat{\text{Cov}}(S_1 O_2, S_2 O_2))}{\sqrt{\hat{\sigma}^2_{O1} \times \hat{\sigma}^2_{O2}}}$
$p \times F \times O \text{ CE} = \dfrac{\hat{\sigma}^2_p + \dfrac{\hat{\sigma}^2_{po}}{n'_o}}{\hat{\sigma}^2_p + \dfrac{\hat{\sigma}^2_{pf}}{n'_f} + \dfrac{\hat{\sigma}^2_{po}}{n'_o} + \dfrac{\hat{\sigma}^2_{pfo,e}}{n'_f n'_o}}$	Forms-based Two Facet CE $= \dfrac{.5 \times (\widehat{\text{Cov}}(F_1 O_1, F_2 O_1) + \widehat{\text{Cov}}(F_1 O_2, F_2 O_2))}{(\hat{\sigma}^2_{F1O1} \times \hat{\sigma}^2_{F2O1} \times \hat{\sigma}^2_{F1O2} \times \hat{\sigma}^2_{F2O2})^{1/4}}$
$p \times I \times O \text{ CS} = \dfrac{\hat{\sigma}^2_p + \dfrac{\hat{\sigma}^2_{pi}}{n'_i}}{\sigma^2_p + \dfrac{\hat{\sigma}^2_{pi}}{n'_i} + \dfrac{\hat{\sigma}^2_{po}}{n'_o} + \dfrac{\hat{\sigma}^2_{pio,e}}{n'_i n'_o}}$	Items-based Two Facet CS = Test-retest $= \dfrac{\widehat{\text{Cov}}(O1, O2)}{\sqrt{\hat{\sigma}^2_{O1} \times \hat{\sigma}^2_{O2}}}$
$p \times S \times O \text{ CS} = \dfrac{\hat{\sigma}^2_p + \dfrac{\hat{\sigma}^2_{ps}}{n'_s}}{\sigma^2_p + \dfrac{\hat{\sigma}^2_{ps}}{n'_s} + \dfrac{\hat{\sigma}^2_{po}}{n'_o} + \dfrac{\hat{\sigma}^2_{pso,e}}{n'_s n'_o}}$	Splits-based Two Facet CS = Test-retest $= \dfrac{\widehat{\text{Cov}}(O1, O2)}{\sqrt{\hat{\sigma}^2_{O1} \times \hat{\sigma}^2_{O2}}}$
$p \times F \times O \text{ CS} = \dfrac{\hat{\sigma}^2_p + \dfrac{\hat{\sigma}^2_{pf}}{n'_f}}{\sigma^2_p + \dfrac{\hat{\sigma}^2_{pf}}{n'_f} + \dfrac{\hat{\sigma}^2_{po}}{n'_o} + \dfrac{\hat{\sigma}^2_{pfo,e}}{n'_o n'_f}}$	Forms-based Two Facet CS $= \dfrac{.5 \times (\widehat{\text{Cov}}(F_1 O_1, F_1 O_2) + \widehat{\text{Cov}}(F_2 O_1, F_2 O_2))}{(\hat{\sigma}^2_{F1O1} \times \hat{\sigma}^2_{F2O1} \times \hat{\sigma}^2_{F1O2} \times \hat{\sigma}^2_{F2O2})^{1/4}}$
$p \times I \times O \text{ CES} = \dfrac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \dfrac{\hat{\sigma}^2_{pi}}{n'_i} + \dfrac{\hat{\sigma}^2_{po}}{n'_o} + \dfrac{\hat{\sigma}^2_{pio,e}}{n'_i n'_o}}$	Items-based CES $= \dfrac{\left(\dfrac{n_{Total}}{n_{Total}\text{-}1}\right) \sum_i^{n_{Total}} \sum_{i \neq j}^{n_{Total}} \widehat{\text{Cov}}(I_i O_1, I_j O_2)}{\sqrt{\hat{\sigma}^2_{O1} \times \hat{\sigma}^2_{O2}}}$

(*Appendices continue*)

Table A2 (*continued*)

G-coefficient formula[a]	Conventional coefficient[b]
$p \times S \times O$ CES $= \dfrac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \dfrac{\hat{\sigma}_{ps}^2}{n'_s} + \dfrac{\hat{\sigma}_{po}^2}{n'_o} + \dfrac{\hat{\sigma}_{pso,e}^2}{n'_s n'_o}}$	Splits-based CES $= \dfrac{2 \times (\widehat{\text{Cov}}(S_1O_1, S_2O_2) + \widehat{\text{Cov}}(S_1O_2, S_2O_1))}{\sqrt{\hat{\sigma}_{O1}^2 \times \hat{\sigma}_{O2}^2}}$
$p \times S \times O$ CES (adjusted) $= .25 \times \left(\dfrac{n_{Total}}{n_{Split\ 1}} \times \dfrac{n_{Total}}{n_{Split\ 2}}\right) \times \left(\dfrac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \dfrac{\hat{\sigma}_{ps}^2}{n'_s} + \dfrac{\hat{\sigma}_{po}^2}{n'_o} + \dfrac{\hat{\sigma}_{pso,e}^2}{n'_s n'_o}}\right)$	Splits-based Two Facet CES (adjusted) $= \dfrac{.5 \times \left(\dfrac{n_{Total}}{n_{Split\ 1}} \times \dfrac{n_{Total}}{n_{Split\ 2}}\right) \times (\widehat{\text{Cov}}(S_1O_1, S_2O_1) + \widehat{\text{Cov}}(S_1O_2, S_2O_2))}{\sqrt{\hat{\sigma}_{O1}^2 \times \hat{\sigma}_{O2}^2}}$
$p \times F \times O$ CES $= \dfrac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \dfrac{\hat{\sigma}_{pf}^2}{n'_f} + \dfrac{\hat{\sigma}_{po}^2}{n'_o} + \dfrac{\hat{\sigma}_{pfo,e}^2}{n'_f n'_o}}$	Forms-based CES $= \dfrac{.5 \times (\widehat{\text{Cov}}(F_1O_1, F_2O_2) + \widehat{\text{Cov}}(F_1O_2, F_2O_1))}{(\hat{\sigma}_{F1O1}^2 \times \hat{\sigma}_{F2O1}^2 \times \hat{\sigma}_{F1O2}^2 \times \hat{\sigma}_{F2O2}^2)^{1/4}}$

Note. n_{Total} = number of items in full scale; $n_{Split\ 1}$ = number of items in split 1; $n_{Split\ 2}$ = number of items in split 2; $S1$ = Split 1 scores; $S2$ = Split 2 scores; $O1$ = Occasion 1 scores; $O2$ = Occasion 2 scores; $F1$ = Form 1 scores; $F2$ = Form 2 scores; S_1O_1 = Split 1 scores on Occasion 1; S_1O_2 = Split 1 scores on Occasion 2; S_2O_1 = Split 2 scores on Occasion 1; S_2O_2: Split 2 scores on Occasion 2; F_1O_1 = Form 1 scores on Occasion 1; F_1O_2 = Form 1 scores on Occasion 2; F_2O_1 = Form 2 scores on Occasion 1; F_2O_2 = Form 2 scores on Occasion 2; I_iO_1 = score on Item i on Occasion 1; I_jO_2 = score on Item j on Occasion 2; I_iO_2 = score on Item i on Occasion 2; I_jO_1 = score on Item j on Occasion 1; CE = Coefficient of Equivalence; CS = Coefficient of Stability; CES = Coefficient of Equivalence and Stability.

[a] For the present examples, n'_o and n'_f both equal 1, n'_s equals 2, and n'_i values range from 8 to 20. [b] For items-based and splits-based analyses, conventional coefficients can be converted to two facet G-coefficients using the conversions that follow:

$$p \times I \times O \text{ or } p \times S \times O \text{ G}-\text{coefficient} = (\text{Conventional coefficient}) \times \frac{\sqrt{\hat{\sigma}_{O1}^2 \times \hat{\sigma}_{O2}^2}}{(\hat{\sigma}_{O1}^2 + \hat{\sigma}_{O2}^2)/2}$$

$$p \times F \times O \text{ G-coefficient} = (\text{Conventional coefficient}) \times \frac{(\hat{\sigma}_{F1O1}^2 \times \hat{\sigma}_{F2O1}^2 \times \hat{\sigma}_{F1O2}^2 \times \hat{\sigma}_{F2O2}^2)^{1/4}}{(\hat{\sigma}_{F1O1}^2 + \hat{\sigma}_{F2O1}^2 + \hat{\sigma}_{F1O2}^2 + \hat{\sigma}_{F2O2}^2)/4}$$

(*Appendices continue*)

Table A3
Formulas for G-Theory Dependability Coefficients for One- and Two-Facet Designs

Global D-coefficients[a]	Cut-score based D-coefficients[a]
$p \times I$	
$\text{Global D-coefficient}_{p\times I} = \dfrac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \dfrac{\hat{\sigma}_{pi,e}^2}{n_i'} + \dfrac{\hat{\sigma}_i^2}{n_i'}}$	$\text{Cut-score based D-coefficient}_{p\times I} = \dfrac{\hat{\sigma}_p^2 + (\overline{Y} - C) - \hat{\sigma}_Y^2}{\hat{\sigma}_p^2 + [(\overline{Y} - C)^2 - \hat{\sigma}_Y^2] + \left[\dfrac{\hat{\sigma}_{pi,e}^2}{n_i'} + \dfrac{\hat{\sigma}_i^2}{n_i'}\right]}$, where $\hat{\sigma}_Y^2 = \dfrac{\hat{\sigma}_p^2}{n_p'} + \dfrac{\hat{\sigma}_{pi,e}^2}{n_p' n_i'} + \dfrac{\hat{\sigma}_i^2}{n_i'}$
$p \times S$	
$\text{Global D-coefficient}_{p\times S} = \dfrac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \dfrac{\hat{\sigma}_{ps,e}^2}{n_s'} + \dfrac{\hat{\sigma}_s^2}{n_s'}}$	$\text{Cut-score based D-coefficient}_{p\times S} = \dfrac{\hat{\sigma}_p^2 + (\overline{Y} - C)^2 - \hat{\sigma}_Y^2}{\hat{\sigma}_p^2 + [(\overline{Y} - C)^2 - \hat{\sigma}_Y^2] + \left[\dfrac{\hat{\sigma}_{ps,e}^2}{n_s'} + \dfrac{\hat{\sigma}_s^2}{n_s'}\right]}$, where $\hat{\sigma}_Y^2 = \dfrac{\hat{\sigma}_p^2}{n_p'} + \dfrac{\hat{\sigma}_{ps,e}^2}{n_p' n_s'} + \dfrac{\hat{\sigma}_s^2}{n_s'}$
$p \times F$	
$\text{Global D-coefficient}_{p\times F} = \dfrac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \dfrac{\hat{\sigma}_{pf,e}^2}{n_f'} + \dfrac{\hat{\sigma}_f^2}{n_f'}}$	$\text{Cut-score based D-coefficient}_{p\times F} = \dfrac{\hat{\sigma}_p^2 + (\overline{Y} - C)^2 - \hat{\sigma}_Y^2}{\hat{\sigma}_p^2 + [(\overline{Y} - C)^2 - \hat{\sigma}_Y^2] + \left[\dfrac{\hat{\sigma}_{pf,e}^2}{n_f'} + \dfrac{\hat{\sigma}_f^2}{n_f'}\right]}$, where $\hat{\sigma}_Y^2 = \dfrac{\hat{\sigma}_p^2}{n_p'} + \dfrac{\hat{\sigma}_{pf,e}^2}{n_p' n_f'} + \dfrac{\hat{\sigma}_f^2}{n_f'}$
$p \times O$	
$\text{Global D-coefficient}_{p\times O} = \dfrac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \dfrac{\hat{\sigma}_{po,e}^2}{n_o'} + \dfrac{\hat{\sigma}_o^2}{n_o'}}$	$\text{Cut-score based D-coefficient}_{p\times O} = \dfrac{\hat{\sigma}_p^2 + (\overline{Y} - C)^2 - \hat{\sigma}_Y^2}{\hat{\sigma}_p^2 + [(\overline{Y} - C)^2 - \hat{\sigma}_Y^2] + \left[\dfrac{\hat{\sigma}_{po,e}^2}{n_o'} + \dfrac{\hat{\sigma}_o^2}{n_o'}\right]}$, where $\hat{\sigma}_Y^2 = \dfrac{\hat{\sigma}_p^2}{n_p'} + \dfrac{\hat{\sigma}_{po,e}^2}{n_p' n_o'} + \dfrac{\hat{\sigma}_o^2}{n_o'}$
$p \times I \times O$	
$\text{Global D-coefficient}_{p\times I\times O} = \dfrac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \dfrac{\hat{\sigma}_{pi}^2}{n_i'} + \dfrac{\hat{\sigma}_{po}^2}{n_o'} + \dfrac{\hat{\sigma}_{pio,e}^2}{n_i' n_o'} + \dfrac{\hat{\sigma}_i^2}{n_i'} + \dfrac{\hat{\sigma}_o^2}{n_o'} + \dfrac{\hat{\sigma}_{io}^2}{n_i' n_o'}}$	$\text{Cut-score based D-coefficient}_{p\times I\times O} = \dfrac{\hat{\sigma}_p^2 + (\overline{Y} - C)^2 - \hat{\sigma}_Y^2}{\hat{\sigma}_p^2 + [(\overline{Y} - C)^2 - \hat{\sigma}_Y^2] + \left(\dfrac{\hat{\sigma}_{pi}^2}{n_i'} + \dfrac{\hat{\sigma}_{po}^2}{n_o'} + \dfrac{\hat{\sigma}_{pio,e}^2}{n_i' n_o'} + \dfrac{\hat{\sigma}_i^2}{n_i'} + \dfrac{\hat{\sigma}_o^2}{n_o'} + \dfrac{\hat{\sigma}_{io}^2}{n_i' n_o'}\right)}$, where $\hat{\sigma}_Y^2 = \dfrac{\hat{\sigma}_p^2}{n_p'} + \dfrac{\hat{\sigma}_{pi}^2}{n_p' n_i'} + \dfrac{\hat{\sigma}_{po}^2}{n_p' n_o'} + \dfrac{\hat{\sigma}_{pio,e}^2}{n_p' n_i' n_o'} + \dfrac{\hat{\sigma}_i^2}{n_i'} + \dfrac{\hat{\sigma}_o^2}{n_o'} + \dfrac{\hat{\sigma}_{io}^2}{n_i' n_o'}$
$p \times S \times O$	
$\text{Global D-coefficient}_{p\times S\times O} = \dfrac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \dfrac{\hat{\sigma}_{ps}^2}{n_s'} + \dfrac{\hat{\sigma}_{po}^2}{n_o'} + \dfrac{\hat{\sigma}_{pso,e}^2}{n_s' n_o'} + \dfrac{\hat{\sigma}_s^2}{n_s'} + \dfrac{\hat{\sigma}_o^2}{n_o'} + \dfrac{\hat{\sigma}_{so}^2}{n_s' n_o'}}$	$\text{Cut-score based D-coefficient}_{p\times S\times O} = \dfrac{\hat{\sigma}_p^2 + (\overline{Y} - C)^2 - \hat{\sigma}_Y^2}{\hat{\sigma}_p^2 + [(\overline{Y} - C)^2 - \hat{\sigma}_Y^2] + \left(\dfrac{\hat{\sigma}_{ps}^2}{n_s'} + \dfrac{\hat{\sigma}_{po}^2}{n_o'} + \dfrac{\hat{\sigma}_{pso,e}^2}{n_s' n_o'} + \dfrac{\hat{\sigma}_s^2}{n_s'} + \dfrac{\hat{\sigma}_o^2}{n_o'} + \dfrac{\hat{\sigma}_{so}^2}{n_s' n_o'}\right)}$, where $\hat{\sigma}_Y^2 = \dfrac{\hat{\sigma}_p^2}{n_p'} + \dfrac{\hat{\sigma}_{ps}^2}{n_p' n_s'} + \dfrac{\hat{\sigma}_{po}^2}{n_p' n_o'} + \dfrac{\hat{\sigma}_{pso,e}^2}{n_p' n_s' n_o'} + \dfrac{\hat{\sigma}_s^2}{n_s'} + \dfrac{\hat{\sigma}_o^2}{n_o'} + \dfrac{\hat{\sigma}_{so}^2}{n_s' n_o'}$
$p \times F \times O$	
$\text{Global D-coefficient}_{p\times F\times O} = \dfrac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \dfrac{\hat{\sigma}_{pf}^2}{n_f'} + \dfrac{\hat{\sigma}_{po}^2}{n_o'} + \dfrac{\hat{\sigma}_{pfo,e}^2}{n_f' n_o'} + \dfrac{\hat{\sigma}_f^2}{n_f'} + \dfrac{\hat{\sigma}_o^2}{n_o'} + \dfrac{\hat{\sigma}_{fo}^2}{n_f' n_o'}}$	$\text{Cut-score based D-coefficient}_{p\times F\times O} = \dfrac{\hat{\sigma}_p^2 + (\overline{Y} - C)^2 - \hat{\sigma}_Y^2}{\hat{\sigma}_p^2 + [(\overline{Y} - C)^2 - \hat{\sigma}_Y^2] + \left(\dfrac{\hat{\sigma}_{pf}^2}{n_f'} + \dfrac{\hat{\sigma}_{po}^2}{n_o'} + \dfrac{\hat{\sigma}_{pfo,e}^2}{n_f' n_o'} + \dfrac{\hat{\sigma}_f^2}{n_f'} + \dfrac{\hat{\sigma}_o^2}{n_o'} + \dfrac{\hat{\sigma}_{fo}^2}{n_f' n_o'}\right)}$, where $\hat{\sigma}_Y^2 = \dfrac{\hat{\sigma}_p^2}{n_p'} + \dfrac{\hat{\sigma}_{pf}^2}{n_p' n_f'} + \dfrac{\hat{\sigma}_{po}^2}{n_p' n_o'} + \dfrac{\hat{\sigma}_{pfo,e}^2}{n_p' n_f' n_o'} + \dfrac{\hat{\sigma}_f^2}{n_f'} + \dfrac{\hat{\sigma}_o^2}{n_o'} + \dfrac{\hat{\sigma}_{fo}^2}{n_f' n_o'}$

Note. $\overline{Y}$ (grand mean) and C (cut-score) in the formulas are on the item ($p \times I$ and $p \times I \times O$), split ($p \times S$, $p \times S \times O$), or form ($p \times F$, $p \times F \times O$) metric. Therefore, for inclusion in formulas, the observed total score values for grand mean and cut-score need to be divided by number of items for the $p \times I$ and $p \times I \times O$ designs and by number of splits for the $p \times S$ and $p \times S \times O$ designs.

[a] For the present examples, n_o' and n_f' both equal 1, n_s' equals 2, n_p' equals 206, and n_i' values range from 8 to 20.

(*Appendices continue*)

Appendix B

Estimating D-Coefficients When Using Uneven Splits

For uneven splits, global and cut-score specific D-coefficients will be underestimated using formulas in Table A3 because estimates of split main effects ($\hat{\sigma}_s^2$) are inflated. A method for obtaining more accurate estimates for D-coefficients under these circumstances is to perform a G-theory analysis using an *items nested in splits* design. For single occasion data, the global and cut-score specific D-coefficients would be as follows for a persons × (Items:Splits) design:

$$\text{Global D-coefficient}_{p\times(I:S)} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_s^2}{n'_s} + \frac{\hat{\sigma}_{i:s}^2}{(n'_{i\ per\ split}\, n'_s)} + \frac{\hat{\sigma}_{ps}^2}{n'_s} + \frac{\hat{\sigma}_{p(i:s),e}^2}{(n'_{i\ per\ split}\, n'_s)}}$$

$$\text{Cut-score based D-coefficient}_{p\times(I:S)} = \frac{\hat{\sigma}_p^2 + (\overline{Y} - C)^2 - \hat{\sigma}_Y^2}{\hat{\sigma}_p^2 + \left[(\overline{Y} - C)^2 - \hat{\sigma}_Y^2\right] + \left(\frac{\hat{\sigma}_s^2}{n'_s} + \frac{\hat{\sigma}_{i:s}^2}{(n'_{i\ per\ split}\, n'_s)} + \frac{\hat{\sigma}_{ps}^2}{n'_s} + \frac{\hat{\sigma}_{p(i:s),e}^2}{(n'_{i\ per\ split}\, n'_s)}\right)}$$

For persons × Occasions × (Items:Splits) designs, the formulas would be

$$\text{Global D-coefficient}_{p\times O\times(I:S)} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{ps}^2}{n'_s} + \frac{\hat{\sigma}_{po}^2}{n'_o} + \frac{\hat{\sigma}_{pos}^2}{n'_s n'_o} + \frac{\hat{\sigma}_{p(i:s)}^2}{(n'_{i\ per\ split}\, n'_s)} + \frac{\hat{\sigma}_{po(i:s),e}^2}{n'_o\,(n'_{i\ per\ split}\, n'_s)} + \frac{\hat{\sigma}_s^2}{n'_s} + \frac{\hat{\sigma}_o^2}{n'_o} + \frac{\hat{\sigma}_{i:s}^2}{(n'_{i\ per\ split}\, n'_s)} + \frac{\hat{\sigma}_{so}^2}{n'_s n'_o} + \frac{\hat{\sigma}_{oi:s}^2}{n'_o(n'_{i\ per\ split}\, n'_s)}}$$

$$\text{Cut-score based D-coefficient}_{p\times O\times(I:S)} = \frac{\hat{\sigma}_p^2 + (\overline{Y} - C)^2 - \hat{\sigma}_Y^2}{\hat{\sigma}_p^2 + \left[(\overline{Y} - C)^2 - \hat{\sigma}_Y^2\right] + \left(\frac{\hat{\sigma}_{ps}^2}{n'_s} + \frac{\hat{\sigma}_{po}^2}{n'_o} + \frac{\hat{\sigma}_{pso}^2}{n'_s n'_o} + \frac{\hat{\sigma}_{p(i:s)}^2}{(n'_{i\ per\ split}\, n'_s)} + \frac{\hat{\sigma}_{po(i:s),e}^2}{n'_o\,(n'_{i\ per\ split}\, n'_s)} + \frac{\hat{\sigma}_s^2}{n'_s} + \frac{\hat{\sigma}_o^2}{n'_o} + \frac{\hat{\sigma}_{i:s}^2}{(n'_{i\ per\ split}\, n'_s)} + \frac{\hat{\sigma}_{so}^2}{n'_s n'_o} + \frac{\hat{\sigma}_{oi:s}^2}{n'_o\,(n'_{i\ per\ split}\, n'_s)}\right)},$$

where $\hat{\sigma}_Y^2 = \frac{\hat{\sigma}_p^2}{n'_p} + \frac{\hat{\sigma}_{ps}^2}{n'_p n'_s} + \frac{\hat{\sigma}_{po}^2}{n'_p n'_o} + \frac{\hat{\sigma}_{pso}^2}{n'_p n'_s n'_o} + \frac{\hat{\sigma}_{p(i:s)}^2}{n'_p\,(n'_{i\ per\ split}\, n'_s)} + \frac{\hat{\sigma}_{po(i:s),e}^2}{n'_p n'_o\,(n'_{i\ per\ split}\, n'_s)} + \frac{\hat{\sigma}_s^2}{n'_s} + \frac{\hat{\sigma}_o^2}{n'_o} + \frac{\hat{\sigma}_{i:s}^2}{(n'_{i\ per\ split}\, n'_s)} + \frac{\hat{\sigma}_{so}^2}{n'_s n'_o} + \frac{\hat{\sigma}_{oi:s}^2}{n'_o\,(n'_{i\ per\ split}\, n'_s)}$

Note that $\overline{Y}$ (grand mean) and C (cut-score) in the formulas are on the item metric. Therefore, for inclusion in formulas, the observed total score values for grand mean and cut-score need to be divided by total number of items. In these formulas, $n'_{i\ per\ split}$ = mean number of items per split, and $(n'_{i\ per\ split} \times n'_s)$ = total number of items. For the present examples, n'_o equals 1, n'_s equals 2, n'_p equals 206, and $(n'_{i\ per\ split} \times n'_s)$ values range from 8 to 20.

Received December 4, 2015
Revision received July 13, 2016
Accepted July 14, 2016 ■