Generalizability Theory: An Explanation, Tutorial, Systematic Review, and Guidelines for Organizational Researchers

Author Note

This study was not preregistered. Primary and simulated data for this article is shared via the link below. We have no conflicts of interest to disclose.

Supplemental material for this article can be found at the following link:

https://osf.io/8v7j2/overview?view_only=ac8d80abd00c4ab5a73afa2f83bd7963

Abstract

Generalizability theory (G theory) was developed by Lee Cronbach and his colleagues as an analytic paradigm for studying measurement structure and reliability. Part of Cronbach's aim with G theory was to furnish the study of psychometrics with concepts that were, previously, almost exclusive to experimental methods. Despite its advantages, G theory is not widely adopted in organizational research. In this paper, we explore reasons why G theory has not been more widely adopted, including a lack of awareness of its capabilities and a sometimes unclearly explained nomenclature. Via tutorial examples, annotated code, and simulated data, we explain G theory, the problems it can be used to address, and how it can be applied. We round out this instructional knowledge with a systematic review of diverse G theory applications from organizational research. Themes are drawn together from our tutorial examples and lessons learned from research applications with a set of G theory guidelines for organizational researchers. Our aim is to facilitate an understanding of G theory, its effective future application, and to highlight its relevance to increasingly complex organizational data sets.

Keywords: generalizability theory, reliability, measurement

Generalizability Theory: An Explanation, Tutorial, Systematic Review, and Guidelines for Organizational Researchers

Generalizability (G) theory is a psychometric framework formally developed by Lee Cronbach and his colleagues for the study of measurement structure and reliability (Cronbach et al., 1972; Cronbach et al., 1963). The core premise of G theory centers on whether it is possible to generalize from a specific measurement (e.g., rater responses on rating items indicating leadership perceptions) to a hypothetical universe (i.e., a large, potentially infinite number) of equivalent measurements that were obtained under measurement conditions that differ from those used originally (e.g., with alternative raters and items deemed to serve an equivalent purpose to the raters and items used originally).

This idea is operationalized in G theory by parsing variance in measurements (e.g., responses on rating items) based on modeled random effects variance components (or *effects*) that represent systematic sources of variance from a measurement procedure (e.g., people being evaluated, items, and raters, Jiang et al., 2020; Marcoulides, 1998). Effects considered to represent contributions to reliable variance (e.g., variability in people being evaluated) are assumed to approximate the grand average of equivalent reliable measurements taken across a universe of different measurement conditions (referred to as *universe score* variance). This concept is analogous to *true score* variance in classical test theory. In contrast, effects considered to contribute *un*reliability (referred to as measurement *error* variance, e.g., idiosyncratic rater variance, specific item variance) are assumed to interfere with the ability to generalize measurements across different conditions. For example, extensive rater- and unique item-related variance present barriers to generalizing across both raters and items to evaluate desired universe score variance. The relative contribution of universe score to error variance is used in G theory to evaluate measurement reliability (Brennan, 2001a). With some restrictions that we discuss below, researchers may exercise judicious decisions about which

sources of variance contribute either to universe score, error variance, or which are irrelevant to scoring objectives.

A key advantage of G theory is its capacity to handle measurement designs of a high level of complexity, such as those involving raters evaluating people on multiple performance dimensions (e.g., communication, teamwork) or psychological constructs (e.g., personality, Vispoel et al., 2024). G theory moreover accommodates bespoke adjustments to reflect different approaches to forming summary scores and different scoring objectives (e.g., the use of cut-scores or a focus on differences between ratees, Webb et al., 2006).

Despite its adaptability, G theory is not typically adopted as the preferred option for the analysis of measurement data in organizational contexts. Relatively simplistic reliability estimates are more commonly applied to organizational data, such as coefficient alpha or intraclass correlations (ICCs, Raykov & Marcoulides, 2019; ten Hove et al., 2024). Several researchers have commented on how the underutilization of G theory presents a missed opportunity for organizational research (e.g., DeShon, 2002; Jackson, Michaelides, Dewberry, & Englert, 2022). Woehr et al. (2012) suggest that declining interest in G theory and its restrictive ANOVA-based assumptions coincided with rising interest in confirmatory factor analysis (CFA) and its more relaxed statistical assumptions. Relatedly, G theory adopts language typically used in experimentation. DeShon (2002) suggests this language renders G theory “inaccessible to the uninitiated” (pp. 194-195). This inaccessibility is unfortunate given Cronbach’s aim of making invaluable experimental concepts available to measurement research (Cronbach et al., 1972).

In addition to a historical basis in ANOVA and the adoption of a specific language, we raise several other interconnected reasons for the underutilization of G theory. These reasons include that (a) organizational researchers and practitioners have had limited exposure to G theory in their psychometric and statistical training (DeShon, 2002), (b) learning resources

for G theory are often statistically dense (e.g., Brennan, 2001a), (c) specialized G theory software is often statistically restrictive and non-user-friendly (e.g., Brennan, 2001b), and (d) misunderstandings arise about how and to which problems G theory should be applied (e.g., Schmidt et al., 2000). These factors likely create a sustained cycle of unfamiliarity that leads to the first four reasons specified above.

In this paper, we directly address these issues in three ways. First, we explain G theory for organizational research using a series of tutorial examples. Annotated R code and simulated data accompany these examples, which progress from simple to complex designs. Our explanations and examples do not avoid the language of G theory because that language is integral to its intended purpose. However, we explain and at times update the language for a contemporary context. Second, we present a systematic review that provides learning opportunities from widely varied organizational G theory applications. Third, by drawing together observations from our tutorial examples and lessons learned from our review, we contribute guidelines for conducting and reporting G theory analyses.

Our aims extend previous work on G theory. Notably, DeShon (2002) and Putka and Hoffman (2014) presented conceptual and technical introductions. Woehr et al. (2012) presented contrasts between G theory and CFA for multitrait-multimethod structures. Outside organizational methods, Vispoel et al. (2018) provide an excellent technical account of G theory focusing on ANOVA-based estimators for personality assessment. Brennan (1992) covers G theory concepts and ANOVA estimators. In contrast, our paper covers G theory concepts and applications via tutorial examples using modern statistical estimation techniques in the organizational context. We contribute a novel applied research perspective via a systematic review and guidelines for reporting G theory analyses.

We note that our paper primarily focuses on univariate rather than multivariate G theory. This is because, as our review reveals, almost all organizational applications of G

theory have so far adopted a univariate basis. In such instances, even if results from multivariate G theory were to be derived, with independent variables they would lead to the same results as those from a univariate analysis. Also, a proper consideration of multivariate G theory requires an extensive discussion of its ancillary complexities that is beyond our scope. Nonetheless, we mention multivariate G theory where relevant as a promising avenue for future applications and we refer the reader to Raykov and Marcoulides (2011) for a more comprehensive introduction to multivariate G theory.

What is G Theory For and Why Does It Matter?

G theory is used to inform on the measurement structure (what does my instrument measure internally?) and reliability (are my measurements consistent and free from measurement error?) in multifaceted measurement designs. Multifaceted measurement designs involve multiple features that vary systematically (e.g., rating items, human raters, different large language models). Relevant to this context, G theory differs from traditional approaches in that it can be used to model multiple sources of reliability and unreliability simultaneously. The G theory perspective provides a more coherent perspective than is possible with piecemeal reliability estimates that are not parsed for the presence of each other and/or that are not necessarily consistent in their definition of reliability. The G theory framework is consistent in how it defines, estimates, and processes measurement model and reliability information and is flexible in that it can accommodate models of almost any level of complexity for almost any measurement intention. Multifaceted measurement designs are commonplace in organizations and so G theory is highly relevant to this context.

Example Measurement Designs

In this section, we discuss four example measurement designs that we draw on to explain G theory and its application. When following our examples, we encourage the reader to peruse the terminology presented and defined in Table 1 and the accompanying R code and

simulated data provided in our supplemental material that tracks each example in turn. Our examples are based on Bayesian estimation (see Appendix A for a brief primer on Bayes methods). Nonetheless, we provide both R code for Bayesian estimation using the brms package (Version 2.23.0, Bürkner, 2018) and Frequentist estimation using the lme4 package (Version 2.0.1, Bates et al., 2015), given these packages are open-source and optimized for models that are often applied in G theory. Our testing revealed that the results of each analysis were similar, regardless of which statistical basis or package was applied.

Four Analytic Steps

Our examples follow four steps, which we state here in full and explain below in relation to each of our example measurement designs:

- Step 1: Identifying measurement design effects
- Step 2: Identifying how effects are configured
- Step 3: Modeling effects (detailed in our online supplement)
- Step 4: Presenting adjusted estimates

Example 1: p × i Design

Refer to the file "Example 1 p × i.R" in our online supplement.

Step 1. The first step in a G theory analysis involves identifying which elements of a measurement design occur systematically and thus can be modeled. In our first example, 100 participants (followers) completed a 3-item scaled measure of leadership perceptions. The effects relevant to this design include participants (p), items (i), and residual error. In this case, residual error is estimated via the interaction (i.e., the multiplicative outcome of more than one effect) between p and i, which, as it is the highest-order effect (i.e., involves all modeled effects), is confounded with undifferentiated error (pi,*e*).

Step 2. This example is referred to as a *crossed* configuration, where all participants completed all items. Put another way, in a crossed design, all levels of p (i.e., all individual

participants) occur with all levels of i (i.e., all individual items). Such configurations can be represented as p × i designs.

Step 3. Variance components analyses are commonly, although not exclusively (e.g., Vispoel et al., 2024), applied in the context of G theory. Variance components analyses have historical roots in factorial ANOVA. Their purpose in G theory is not as a traditional assessment of statistical significance, but rather as an evaluation of the size of each effect as estimated by partitioning variance in mean squares. We present traditional ANOVA formulas in Table 2 to facilitate a conceptual understanding of this estimation process and a toy analysis worksheet in our supplemental material intended to show the historical ANOVA-based mechanics underlying G theory analyses (see the file "Toy G&D Study.xlsx"). Relative to an ANOVA basis, modern approaches towards the estimation of variance components offer a greater degree of flexibility over handling such issues as data sparsity and unbalanced data configurations (see Table 2). Our supplemental code for the p × i design details how this analysis is conducted.

Step 4. A G study and a Decision (D) study for Example 1 are shown in Table 3. A G study presents the variance components estimates described above, usually in their raw form and after adjustments have been applied (e.g., for aggregation across items to form an overall average summary score). A D study, in contrast, is concerned with applying the adjusted (see below) variance components identified in a G study to the estimation of reliability. Table 3 shows variance component estimates and 95% credible intervals (CrIs) for p, i, and pi,*e*, along with proportions of variance explained for each. Total variance explained can be calculated by summing relevant variances in the Est (estimate) column and computing proportions associated with each individual effect (e.g., p explains 64% of total variance).

Unlike many other methods, G theory allows for estimates to be adjusted to reflect the impact of aggregation. Aggregation refers to the common practice where observations are

summarized (e.g., where individual ratings are averaged) and those summaries are applied in research or to guide decisions. To the right of Table 3, variance components appear that are adjusted to reflect aggregation across items to form overall (O) scores for two scoring perspectives: relative and absolute decisions. First, the *relative decisions* perspective (labelled O-rel) is presented, which is used when research focuses on individual differences. Here, effects are retained for presentation that impact the relative positioning of whomever or whatever is being evaluated (referred to as the *object* or *focus of measurement*[1], which in this case constitutes participants). This includes any effect that involves p, either as a main effect (i.e., p) or an interaction (i.e., pi,*e*). On this point, an advantage of G theory is that even though all potential effects are modeled, effects may be omitted from presentation where they are not relevant to the focus of an analysis. This is possible because variance components are estimated such that they are independent of each other (Searle et al., 1992). Second, the *absolute decisions* (O-abs) perspective is presented, which is relevant to a research interest in comparing summary scores to set criteria (e.g., a pass/fail criterion). In absolute decisions, all effects are relevant to an indication of reliability, including p, i, and pi,*e*.

The effects in Table 3 are adjusted arithmetically to reflect the impact of aggregation. In this case, because the aim is to average across 3 items, all effects concerning i are divided by 3 ($n_i = 3$). It also possible to divide any i-related effects in Table 3 (i.e., i and pi,*e*) by integers other than the observed $n_i = 3$ to provide "what if" type D study projections (Shavelson & Webb, 1991). This approach, which is analogous to the Spearman-Brown prophecy formula, can be used to estimate how many items are required to reach admissible reliability (see below) or how the item number can be truncated without detrimentally

[1] The traditional term is *object of measurement.* We suggest an update in our term *focus of measurement* given that the focus is often people.

impacting reliability. Unlike the Spearman-Brown formula, such projections in G theory can be applied to multiple effects (e.g., item- and rater-related effects).

In the lower portion of Table 3 appears the D study elements of the analysis. The D study in this case includes reliability estimates for relative decisions (termed *G coefficients*) and for absolute decisions (termed *Phi coefficients*). See Appendix B and C for full G and Phi coefficient equations used in our examples. To compute each coefficient, it is necessary to take the adjusted estimates in the O-rel and O-abs columns in Table 3 and to classify them as either contributions to reliable variance (referred to as *universe score variance*, defined in this case solely by the p effect as the representation of construct variance) or unreliable variance (referred to as *error variance*, defined here by effects other than p). The individual variance components that represent error variance are referred to as *facets*. Reliability is estimated by computing the relative proportion of universe score to error variance (formulas are provided in brief in the Table 3 note and in full in Appendix B and C). In the current example, the resulting G coefficient (.88, CrI [.83, .91]) is almost identical to coefficient alpha (which we also present in Table 3 for comparison) because it approximates the same statistic. The contrasting Phi coefficient, with its additional consideration of the item main effect facet, is marginally lower at .84, CrI [.56, .89]. Although this presentation refers to a simple p × i design, the same basic principles above apply to any design, regardless of its complexity.

We note here that within our scope; we focus on G coefficients and general-purpose Phi coefficients. However, other coefficients are available within the G theory framework, including Phi (lambda) or $\phi(\lambda)$, which acknowledges specific cut-score criteria. Formulas for Phi (lambda) can be found in Raykov and Marcoulides (2011, section 9.4).

Example 2: Unbalanced p × (i:d) design

Refer to the file “Example 2 p × id.R” in our online supplement.

Many measurement procedures involve not only a set of items grouped into a single construct as in the case above, but item subsets grouped into different constructs. This is referred to as a *nested* configuration, where unique sets of items are grouped or nested in each of several constructs. Nested configurations are often relevant to applications of CFA and its associated McDonald's omega reliability estimate (e.g., Flora, 2020). Table 4 presents a nested design where subsets of items are nested in each of 5 leadership dimensions (d, e.g., business acumen, influence). In G theory notation, a colon denotes a level of nesting, where the nesting of i in d is presented as i:d. Participants in this case completed all items on all dimensions. Thus, the overall design in Table 4 is partially nested, where participants are crossed with items and items are nested in dimensions, $p \times (i:d)$[2].

Because we focus on univariate G theory, the perspective for scoring different dimensions will provide estimates equivalent to the *average expected outcome for any given dimension*. Multivariate G theory, in contrast, provides a psychometric perspective on how specific dimensions differ from each other (see Raykov & Marcoulides, 2011).

Steps 1, 2, and 3. On Step 1, 100 participants rated 24 items relating to 5 dimensions. On Step 2, effects are configured such that participants are crossed with items. An unbalanced number of items (3, 4, 7, 4, and 6 items) are nested in each dimension. This design yields the main effects p, d, and i:d. The fact that i is nested in d means that i is confounded with d and thus no interactions can involve the crossing of i with d. Thus, here it is only possible to estimate the interactions pd and the residual pi:d,*e*. On Step 3, the effects in Table 4 can be modeled with a variance components analysis as per our $p \times (i:d)$ online supplement.

Step 4. On Step 4, results are presented using the same principles as those in the previous example. However, added in Table 4 are the effects d and i:d as well as two relevant

[2] This notation is modified from that traditionally applied in G theory. Here, parentheses or brackets around nested effects are intended to add clarity.

scoring perspectives, including those for (a) a given dimension score and (b) overall scores averaged across different dimensions. Relative decision adjustments are presented for both dimension and overall scores. However, absolute decision adjustments are presented only for overall scores because, in principle, scores could differ by dimension and information relating to specific dimensions is unavailable in univariate G theory.

To adjust for aggregation to a given dimension score in an unbalanced design, any effects involving the facet i are divided by the harmonic mean number of items nested in each dimension (Brennan, 2001a). This adjustment is taken as a conservative approximation of the number of items nested in each dimension ($n_{i:d}$)[3]. In contrast, aggregation occurs for overall scores across an approximation of all items and dimensions. Here, any effect involving dimensions is divided by the number of dimensions (n_d) and any effect involving items is divided by the harmonic mean number of items per dimension multiplied by the number of dimensions, $n_{i:d}n_d$, to reflect the total item n. Rather than using the total item n, the harmonic mean is used in this formula to maintain consistency with the other calculations in Table 4.

Turning to the D study component of the analysis, universe score variance for dimension scores includes the effect p (reflecting general dimension variance) + pd (reflecting unique dimension variance). For overall scores, universe score variance is defined solely by p because all other effects involve aggregation and therefore contribute disturbances to the overall score aggregate. For example, pd is divided by n_d because the intent is to aggregate across dimensions to form an overall score. Thus, any variance in pd interferes with the ability to form an overall aggregate. Table 4 shows a high G coefficient for dimension scores, a similar G coefficient for overall scores, and a slightly lower value for the Phi coefficient, given that it is impacted by a greater number of facets.

[3] We modify this from traditional G theory notation where $n_{i:d}$ is often specified as n_i. However, the latter could be misinterpreted as the total number of items.

Shown for comparison, composite omega in Table 4 is slightly higher than its closest G theory analogue, the G coefficient for dimension scores. This difference stems from contrasting psychometric assumptions. Factor intercorrelations and factor loadings are freely estimated in omega's congeneric CFA basis. By contrast, an overall unique dimension effect is estimated in G theory along with a general effect that subsumes effect covariances. Moreover, unlike CFA, the variance components analyses often used in G theory assume essential tau equivalence (i.e., that rating means are free to vary, but items are considered to contribute equally to an aggregated score, Raykov & Marcoulides, 2006)[4]. While these restrictions lead to more conservative lower-bound reliability estimates (McNeish, 2018), there are notable advantages in that tau equivalent estimates (a) accommodate greater complexity than is possible with congeneric models (e.g., multitrait-multimethod structures are more likely to converge with a variance components analysis than with CFA, Woehr et al., 2012) and (b) better match the common practice of creating aggregate scores based on averaged observations, which are not optimized around specific factor loadings.

Example 3: (p:l) × i Design

Refer to the file "Example 3 pl × i.R" in our online supplement.

In some research contexts (e.g., ratings of leaders by followers), measurement data is collected at two levels. Table 5 presents an example design of this type where psychometric information is presented at both the follower level 1 and leader level 2. The design in Table 5 is a multilevel version of the p × i configuration presented earlier, where participant followers are nested in leaders (p:l) who rate all (and are thus crossed with) items. The overall design in Table 5 can therefore be represented by (p:l) × i.

Steps 1, 2, and 3. On Step 1, 3,000 participant followers (p) rated perceptions of their leader (l, n = 100) on 3 item responses. On Step 2, effects are configured such that

[4] However, essential tau equivalence is not inherently assumed in G theory.

participants are nested in leaders (p:l, with a balanced 30 participants nested in each leader). Participants completed all items and so participants are crossed with items. This design yields the main effects p, p:l, and i. Given partial nesting, it is only possible to estimate the interactions li and the residual effect pi:l,*e*. On Step 3, the effects p, p:l, i, li, and pi:l,*e* can be modeled with a variance components analysis as per our supplemental (p:l) × i code.

Step 4. The results presented in Table 5 are similar to those presented above for our other examples. However, a key difference in Table 5 is that with a multilevel (p:l) × i model, it is possible to specify perspectives for more than one focus of measurement, in this case for level 1 participant followers and for level 2 leaders. For the follower perspective, effects specifically relating to leaders (l and li) are ignored. For relative level 1 decisions, this leaves p:l (the focus of measurement in this case) and the residual estimate adjusted for aggregation across 3 items, pi:l,*e*/n_{i}. For level 1 absolute decisions, the i facet is included because it could impact comparisons with set scoring criteria. For the contrasting level 2 leader perspective, leader-related effects are retained and the focus of measurement shifts from p:l to the l effect. Interest here lies in aggregating across followers within each leader's group. Thus, effects involving the p:l effect are divided by the $n_{\text{p:l}}$ (i.e., 30). It is reasonable for researchers to set this divisor for p:l to one if the aim is to estimate reliability expected for a given follower. Note also that in aggregation to overall scores for effects implicating both p:l and i effects (i.e., in pi:l,*e* in Table 5), the divisor is the product of $n_{\text{p:l}}n_{\text{i}}$.

D study results are presented in the lower portion of Table 5. Only overall scores are relevant to this design because there are no sub-dimensions. G and Phi coefficients for level 1 are lower than for level 2, indicating differences in cross-level reliability. Phi coefficients are generally lower than the G coefficients, given that more facets could impact unreliability in absolute decisions (see the O1-abs and O2-abs columns in Table 5). Also shown for comparison in Table 5 are multilevel omega coefficients. Given the absence of a

multidimensional structure (see discussion above), these comparator estimates trace closely the corresponding G coefficients for this analysis.

Example 4: Balanced (p:l) × (i:d) design

Refer to the file "pl × id.R" in our online supplement.

Our final example adds a multidimensional structure to the multilevel design above, where the intent is to evaluate follower perceptions of leaders based on distinct dimensions. As with the previous example, the multilevel structure means that it is once again possible to present focus-of-measurement perspectives relating to followers (level 1) and leaders (level 2). To simplify its presentation, this analysis is split across two tables. Table 6 deals with expected reliability for aggregation to a dimension scores. Table 7 deals with aggregation across dimensions to form overall scores. In this example, participant followers were nested in leaders (p:l), who they rated on items that were nested in different dimensions. This can be summarized as a (p:l) × (i:d) design.

Steps 1, 2, and 3. On Step 1, 3,000 participant followers (p) rated perceptions of their leader (l, of whom there were 100) on 15 item responses corresponding to 5 dimensions. On Step 2, effects are configured such that participants are nested in leaders (p:l, with a balanced 30 participants nested in each leader). Items are nested in dimensions (i:d, with a balanced 3 items nested in each of 5 dimensions). All participants rated all items, so participants and items are crossed. This design yields the main effects l, p:l, d, and i:d. Considering its partially nested configuration, the design produces the interactions ld, li:d, pd:l, and the residual effect pi:ld,*e*. On Step 3, these effects can be modeled with a variance components analysis, which can be found in our supplemental material for the (p:l) × (i:d) design.

Step 4. The results in Tables 6 and 7 are similar to the previous multilevel example in that they acknowledge level 1 and 2 perspectives. However, in this instance, the analysis integrates a multidimensional structure. In Table 6, only relative decision information is

presented because of the presence of multiple dimensions (as previously discussed). The effects d and i:d in Table 6 are modeled but are ignored for presentation purposes because they are irrelevant to relative decisions. For level 1, effects specifically concerning leaders are omitted from presentation because the focus of measurement is followers. This leaves the effects p:l (the general variance effect and focus of measurement), pd:l (unique dimension variance), and the residual effect, pi:ld,*e*. The residual estimate involves item variance and is therefore divided by $n_{i:d}$. For level 2, leader effects are retained because leaders are the focus of measurement. This focus retains for presentation all effects that include the l variance component and that are therefore pertinent to relative decisions.

Interest in Table 6 is not only in aggregating across items for each dimension, but also, at level 2, in aggregating across followers in leader groups. Consequently, any level 2 effects involving p:l are divided by $n_{p:l}$ (i.e., 30) and, across both levels, any effects involving items are divided by $n_{i:d}$ (i.e., 3). As with previous examples, effects can involve multiple divisors. For example, the level 2 pi:ld,*e* effect is divided by the product of $n_{p:l}n_{i:d}$ because it involves aggregation across both follower (p:l) and item (i:d) effects.

The D study information for Table 6 defines universe score variance as the addition of general (level 1 = p:l, level 2 = l) + unique dimension (level 1 = pd:l, level 2 = ld) variance. Remaining effects specified in the D1-rel (dimension level 1 relative decision) and D2-rel (dimension level 2 relative decision) columns are specified as contributions to error variance. As in the previous example, what is considered a *facet* depends on the focus of measurement. For example, referring to the note in Table 6 for level 1 G and Phi coefficients, the p:l effect contributes to universe score variance. Whereas at level 2, p:l contributes to error variance.

It can be observed in Table 6 that the G coefficient for level 1, .82, CrI [.81, 82], is lower than that for level 2, .93, CrI [.92, .95]. Perhaps the closest traditional analogue of these G coefficients, given the focus on dimensions, is the multilevel version of the composite

omega coefficient. Due to assumption differences described in the previous multidimensional analysis, the omega coefficients in Table 6 are somewhat higher than the G coefficients.

The results in Table 7 are similar to those in Table 6, but with two key differences. The first difference is that the Table 7 results are adjusted for aggregation to overall scores rather than to results for a given dimension. This can be observed in the formulas across levels, where effects concerning d are divided by the n_d and effects concerning i are divided by all items (i.e., $n_{i:d}n_d$). As in the previous example, an additional adjustment is applied in level 2 for aggregation across participants nested in leaders ($n_{p:l}$). Level 1 G and Phi coefficients in Table 7 are lower than level 2 coefficients.

The second difference is in the comparator estimate presented in Table 7. Given the focus here on overall scores, a multilevel variant of ICC2 is presented for comparison adjusted for item and follower *n* (based on Shrout & Fleiss, 1979). ICC2 adopts a different analytic approach to that used in a G coefficient for a (p:l) × (i:d) design, where it partitions fewer sources of variance. For level 1 in Table 7, ICC2 partitions two effects (p and the residual) and for level 2, ICC2 partitions three effects (l, p:l, and the residual). This contrasts against the eight effects partitioned in the G study basis for the G coefficients in Table 7. The implication is that unmodeled sources of variance in ICC2 may unintentionally contribute to either reliable or unreliable variance due to statistical conflation.

The examples above are intended to provide a learning resource and a demonstration of G theory. G theory makes explicit which elements of a measurement design contribute to reliability and unreliability, offers a flexible approach towards acknowledging aggregation, and can be readily applied to multilevel structures. Multiple, statistically parsed perspectives on reliability are combined in G theory. For instance, it is possible to simultaneously model item-related variance (cf. coefficient alpha) and idiosyncratic rater-related variance (cf. ICC2) whilst parsing relations between these two estimates. G theory thus presents a unified

variance-partitioning framework that subsumes different reliability perspectives. It can handle simple or complex measurement designs (see Brennan, 2001a, p. 274). As we explore below, the measurement designs applied in research and practice often present nuances that require a tailored approach to which G theory is well suited.

A Systematic Review of G Theory Applications in Organizational Research

The examples above are for instructional purposes and do not address some of the nuances of data from measurement designs in organizational research. In this section, we systematically review G theory applications to reveal the varied contexts in which it has been applied. Our intention is to demonstrate the versatility of G theory and, in the spirit of continuous improvement, to learn lessons that may assist in refining future applications of the approach as summarized in a set of dedicated G theory guidelines.

Review Protocol

Following the PRISMA systematic review protocol (Page et al., 2021), we conducted a literature search that included the Web of Science, APA PsycInfo, and PsycArticles via Ovid, and ABI/Inform Complete. We used a range of keywords in our search to help ensure comprehensiveness and we applied our prior knowledge that not all authors specified their application of G theory in keywords or abstracts. Post review, we found that only 6/28 studies we identified adopted the keyword “generalizability theory” in their publications. Examples of keywords we used included "G-theory", "Variance components", "generalizability coefficients", "generalizability theory", "G-study", "random effects", "mixed effects", "notion of exchangeability"[5]. Subject selection and keywords varied across database subject categorizations, but we aimed to ensure comprehensive coverage by including the terms *Management, Applied Psychology, Operations Research*, and *Entrepreneurship*. Our inclusion criteria involved empirical articles published from the year 2000 onwards. We

[5] Searches included and excluded hyphens and included UK and US spelling where relevant.

excluded publications from education, counselling, engineering, and marketing because our paper is aimed at reviewing applications of and recent advancements in G theory for organizational research. Detailed information about our search protocol can be found in our PRISMA flow chart (see Figure S1 in our online supplement).

In reporting our findings, we emphasize general patterns observed across studies. We deliberately adopt this approach to present a constructive, solution-oriented perspective rather than one directed towards individual criticism. However, we balance these aims against those concerning transparency. To this end, our review database, which is available in our supplemental material, identifies individual contributing studies.

Systematic Review Findings

Table 8 summarizes general features of the 28 papers we identified where G theory was applied within our set parameters. These studies were based on samples from Europe (k = 8), the USA (k = 7), Asia (k = 3), and New Zealand (k = 1), with nine unreported geographical origins. The largest group of these studies (k = 11) involved singular papers published in a range of different journals. Most studies adopted a Frequentist basis (k = 22). Researchers applied a range of different software options for G theory analyses, including specialized (e.g., urGENOVA) and generic packages.

Table 9 shows frequencies for G study content topics and reveals wide ranging applications. Assessment center (AC) research applied G theory with the highest frequency (k = 10), with smaller frequencies ($k \leq 3$) in the study of competencies, performance ratings, 360-degree ratings, work simulations, and leadership studies. The remaining k = 7 studies concerned a range of different topics, including job analysis, corporate reputation, creativity via evaluations of floral arrangements, and body posture evaluations.

Despite the broad range of topics to which G theory has been applied, common patterns emerged in our review about research design, analytical procedures, and information

availability. The research design issues we identified raise questions about the degree of uncertainty in reported estimates. The analytical patterns we identified are relevant to considering clarity and precision in reported outcomes. The information availability patterns we identified are critical for promoting clarity, open science, and reproducibility. Below, we cover each of these issues in turn as they relate to our findings and combine this knowledge with our earlier example applications to build G theory guidelines for researchers.

Research Design

Small-*n* and the Focus of Measurement. The varied focus of measurement examples from our review included people (e.g., job candidates and employees), videos of behavioral samples, body postures, floral designs, and different corporations. As shown in Table 9, a total of eight studies (i.e., 28 studies total – 20 where focus of measurement $n \geq 30 = 8$) reported an n for their focus of measurement lower than 30 (ranging from $n = 4$ to $n = 20$). A small n for an effect will typically not raise problems when the effect in question is ancillary to the aims of a research study or where the primary interest lies in how the small-n effect interacts with other effects. For example, an AC might include an $n = 4$ for the AC exercise main effect. But, if the focus is on relative decisions, the exercise main effect is largely irrelevant. The interaction between participants and exercises will likely be of interest and will often yield a much higher n than that for the exercise main effect (e.g., for 100 participants, 4 exercises and a crossed design, the interaction $n = 400$).

Any effect relevant to the researcher's aims in a G study requires an n of a size sufficient to approximate a data distribution. The effect representing the focus of measurement is of particular concern in this respect, given that it is almost always a main effect and not an interaction. In accordance with Central Limit Theorem and findings from Monte Carlo simulation studies, an n of at least 30 may provide a starting point for distributional approximation (see Austin, 2010; Howell, 2007, pp. 170–171). However, in

many cases a sample size greatly exceeding 30 will be required to minimize uncertainty (e.g., CrIs) around point estimates. A small-*n* focus of measurement is thus problematic in a G study for two main reasons. First, it will lead to a wide margin of error around a random effect point estimate. This will create interpretational difficulties for G study findings. Second, it will either define all or a substantial portion of universe score variance, thereby creating a high degree of uncertainty in a reliability coefficient. To express this concern in familiar terms, the focus of measurement in coefficient alpha is usually represented by participants. For the same reasons specified above for reliability coefficients in G theory, it would not be meaningful to estimate alpha with a very small group of people (e.g., $n = 4$).

In most (7/8) of the small *n* cases we observed, the focus of measurement was a stimulus other than a group of participants (e.g., recorded stimuli). In the affected seven studies, the effect for the focus of measurement typically reflected a small *n* while other effects in the measurement design (e.g., raters) often reflected more substantial sample sizes. It is therefore possible that confusion arises about which effect(s) require sufficiently large sample sizes in G theory applications. We suggest that researchers form a clear idea about what their focus of measurement is and ensure a sufficiently large *n* for that effect. Our Guideline 1A suggests a minimum *n* of 30 for effects (see Table 10). However, CrIs (or confidence intervals, CIs) will suggest a high degree of uncertainty in cases where an *n* is likely too low. To provide clarity, we recommend that sample sizes *per effect* are reported for all G studies (see Guideline 1B, Table 10).

Data Deletion. Another research design concern we observed was that in three studies, large portions of data were discarded to overcome sparsity (i.e., missing data) and to achieve a fully crossed design. Sparsity can arise because of the way in which a measurement procedure is designed (e.g., where leadership attributes are not evaluated for some but not all leaders in an organization). In the factorial analyses applied in G theory, fully crossed designs

allow for the estimation of a larger number of separate effects. Also with some Frequentist techniques, it is not possible to estimate random effects when a data set is sparse (e.g., in analogous-ANOVA procedures, Brennan, 2001b). Table 8 shows at least 7 studies that employed estimates of this type. Statistical techniques are now widely available that can handle sparse data sets (e.g., Bayesian and restricted maximum likelihood, REML, estimation) and eliminate the requirement to delete data to avoid sparsity. Data deletion for these purposes can lead to undesirable consequences, including that the data subset retained for analysis is not representative of the complete data set. We recommend using estimators that can handle sparse data (see Guideline 1C, Table 10).

Multidimensional Designs. Most of the studies in our review (k = 23) were developed with the intent of evaluating multiple constructs. Every study in our review either (a) reported, using univariate G theory, an interaction between the focus of measurement and the construct effect to indicate variability unique to specific constructs (k = 18) or (b) estimated reliability separately with different G studies for each construct (k = 3, with an additional two studies where this information was unclear).

The interaction approach described above based on univariate G theory informs meaningfully on expected outcomes for a given construct in a multidimensional design. The contrasting approach based on estimating reliability with separate G studies per construct conflates covarying features of the measurement design (e.g., construct covariances). Estimating separate G studies per construct is analogous to modeling a single latent variable CFA per construct. Doing so will fail to account for between-construct correlations and will likely lead to distorted variance estimates. Thus, we suggest avoiding the latter approach. If separate reliability estimates are required for different constructs evaluated within the same procedure, we recommend multivariate G theory (see Guidelines 1D and 1E, Table 10).

Analytical Procedures

Estimation Basis. Our review revealed that G theory had been applied with either Frequentist (n = 22) or Bayesian (n = 6) estimation procedures. There are two possible reasons why Bayesian inference has not been as popular in G theory applications. First, Bayesian G theory was first applied in the organizational literature only around a decade ago (Jackson et al., 2016; LoPilato et al., 2015). Second, accessible software for analysis using Bayesian methods has only become available in fairly recent history with the development of the brms R package (Bürkner, 2017, 2018).

Bayesian analysis carries distinct advantages over Frequentist approaches, particularly in the context of G theory. One of the disadvantages of Frequentist approaches (e.g., analogous-ANOVA, REML) is that they can return negative or zero-fenced variance estimates (Shavelson & Webb, 2005). Negative variances are nonsensical because, by definition, variances are greater than zero. Null variances are possible, although unlikely. Negative and zero-fenced outcomes are avoided in a Bayesian frame because researchers can specify weakly informative priors that prevent such outcomes (Kruschke, 2014).

Another advantage associated with applying Bayesian analyses in G theory is that CrIs are available for all estimated model parameters: including for variance estimates and reliability coefficients. Frequentist CIs, on the other hand, are notoriously difficult to estimate for random effects, are typically unavailable for G or Phi coefficients, and are not available for all Frequentist estimators (see Brennan, 2001a; Searle et al., 1992). We therefore recommend the use of Bayesian over Frequentist estimation for G theory applications (see Guideline 2A, Table 10). However, where researchers wish to apply Frequentist methods, we recommend REML estimation because of its ability to accommodate sparse data and to estimate CIs for random effects (Marcoulides, 1990).

We recognize that Bayesian estimation is generally less common in the organizational literature and that some researchers may not be familiar with how to report analyses based on

this approach. Kruschke (2021), in his open access article, provides clear and accessible guidelines for reporting Bayesian analyses. As a minimum for Bayesian G theory analyses with reference to Kruschke's guidelines (see Table 1 from Kruschke, p. 1284), we recommend reporting points 1A (explain dependent and independent variables), 1C (justify priors), 2A (report specifics about software and packages used), B (report model convergence), C (report effective sample sizes for Monte Carlo Markov Chains [MCMC], see Appendix A), D (report limits of using non-MCMC procedures), 3B (summarize posterior), and 6A (explain software and apply non-proprietary software if possible), B (provide software version information). As needed or optionally, we recommend reporting points 1B (explain likelihood function and model parameters), D (provide model specification), 5A (for weakly informed priors, show that other such priors return similar results), B (compare informed priors with weakly or differently informed priors), or C (sensitivity test default priors), 6C (provide analysis script), D (explain analysis script), G (provide long running MCMC chains), and/or H (provide seed numbers for reproducibility).

Of the studies in our review that applied Bayesian inference, it was unclear whether researchers had based their reported estimates on squared standard deviations from a summary output or on squared values estimated directly from posterior distributions. The former approach results in a form of underestimation known as Jensen's inequality, where a squared summary standard deviation from a posterior distribution does not equal a summary based on posterior estimates that are squared prior to being summarized (Rudin, 1987). To this end, we recommend first squaring values from the posterior and then reporting summaries based on these pre-squared values. To provide further guidance, we apply this approach in the R code provided for our examples (see Guideline 2B, Table 10).

Credible and Confidence Intervals. Our review revealed that very few studies estimated either CrIs or CIs. Only 7/28 studies reported CrIs or CIs for their G study random

effects. Of the 25 studies that reported G or Phi coefficients, only two reported CrIs or CIs. For Frequentist analyses, CIs are only available for some software packages and estimator types (e.g., CIs for random effects analyses are not currently available in IBM SPSS and are not generated for MIVQUE estimators). In software where CIs are available (e.g., in lme4), these cannot be generated for negative or zero-fenced estimates. However, in Bayesian estimation, CrIs are available for all random effects, G, and Phi coefficients.

When reporting statistical outcomes, it is generally recommended that CrIs or CIs should be reported for effect size estimates (e.g., see journal article reporting standards, JARS-Quant Table 1 in American Psychological Association, 2025). This recommendation is particularly important in a G study, where a small effect *n* will drive a wide margin of error. This phenomenon is often referred to as the *Achilles' Heel* critique of G theory (Brennan, 2001a). Without an estimate of the margin of error around a point estimate, it is impossible to determine the impact of the Achilles' Heel critique. For this reason, we recommend that at a minimum, CrIs or CIs are reported for random effects (see Guideline 2C, Table 10). We also suggest similar reporting for G or Phi coefficients. However, only Bayesian techniques currently provide a straightforward approach towards CrI estimation for G or Phi coefficients by allowing researchers to sample from posterior distributions they will generate anyway for their associated G study (see Guideline 2D, Table 10).

Measurement Ill-Structure. We discuss three final points relating to analytic procedures involving measurement ill-structure, adjustments for balanced versus unbalanced designs, and reliability projection via D studies. We found that 11/28 studies in our review included measurement designs that presented candidates for adjustment due to *measurement ill-structure*. Measurement ill-structure usually describes a configuration where raters are allocated to ratees in such a manner that lies between the extremes of a fully-crossed and a perfectly nested configuration (i.e., raters and ratees overlap to some degree). To elaborate,

Putka et al. (2008) developed the *q*-multiplier statistic that adjusts for ill-structure. The *q*-multiplier summarizes the degree of overlap in the allocation of raters (e.g., in a performance evaluation) to ratees (e.g., employees who are being evaluated). A *q* of zero indicates complete overlap, where all raters evaluate all ratees (i.e., a crossed design). The value of *q* increases with diminishing rater-ratee overlap to an upper limit of $1/\hat{k}$ ($\hat{k}$ = the harmonic mean number of raters per ratee). At its upper limit, there is no overlap and each rater evaluates a specific group of ratees (i.e., a nested configuration). The *q*-multiplier was applied in most G studies in our review where it was required (8/11). We recommend continued application of the *q*-multiplier where relevant (see Guideline 2E, Table 10), and we note that Putka et al. provide detailed guidance on its application.

Aggregation and Unbalanced Designs. Designs that are unbalanced with respect to nesting (e.g., where there are unequal numbers of items per dimension) are common in practice. For aggregation adjustments in univariate G theory, unbalanced designs require an indication of the central tendency of levels (e.g., employees) that are nested in some other effect (e.g., teams). The usual recommendation in the G theory literature is to apply the harmonic mean number of nested levels as an indication of central tendency for an unbalanced design (Brennan, 2001a). This is because the harmonic mean is dominated by the smallest number of nested levels and thus tends to be more conservative as a divisor than the arithmetic mean or other indicators of central tendency.

In our review, there were eight studies where it was necessary to estimate a central tendency for unbalanced nested effects. Four of these studies applied the harmonic mean; one study applied the arithmetic mean; and in three studies the choice of central tendency estimate was unclear. We recommend applying and reporting the harmonic mean as an indication of central tendency when adjusting for aggregation for unbalanced nested effects in univariate G studies (see Guideline 2F, Table 10).

D Studies Applied to Projections. When applied to projections, D studies can serve two key purposes. First, they can be used in forecasting for measurement models. For example, D studies can be used when researchers wish to know how many raters are required for a leadership style evaluation to reach an acceptable level of reliability. This approach can be informative for research or for achieving greater resource efficiency (e.g., minimizing the number of raters and items required to reach acceptable reliability). Second, D studies can assist where researchers wish to compare like-with-like conditions across measurement designs. For example, consider an example where two evaluations of leadership style return different reliability estimates. However, only four rating items are included in one evaluation whereas the other evaluation includes 12 rating items. D studies allow researchers to project to a standard number of items across both studies to foster a more equitable comparison. This approach can be applied to any number of measurement conditions (e.g., rating items and raters). We found that D studies for projecting estimates had only been applied in 8/25 studies (i.e., out of the 25 that applied G coefficients, see Table 9). We suggest that researchers consider applying projective D studies, particularly if they facilitate comparisons between measurement conditions (see Guideline 2G, Table 10).

Presentation of Results

Direct versus Illustrative G or Phi Coefficients. Our review revealed a distinction between G theory reliability coefficients intended to directly represent score reliability (which we refer to as *direct coefficients*, i.e., coefficients that directly estimate score reliability) from those intended to illustrate theoretical points (which we refer to as *illustrative coefficients*, i.e., coefficients intended to support an argument, e.g., the extent to which AC exercise variance could hypothetically contribute to AC dimension score reliability). As their universe score variance composition, direct coefficients only include effects that are relevant to score

reliability. By contrast, indirect coefficients include in their universe composition effects that are not directly relevant to score reliability.

To provide an example with a G coefficient, consider the level 2 (p:l) × (i:d) design in Table 6. A *direct* univariate G coefficient for dimension scores is defined here by the addition of the effects l + ld to reflect general and unique leadership dimension effects respectively. A contrasting *illustrative* G coefficient could add to this definition of universe score variance the li:d/$n_{i:d}$ effect to offer hypothetical suggestions about the impact of item-related variance in reliability estimation. However, this illustrative universe score variance definition bears no relevance to the reliability of leadership dimension scores. This is because it will be necessary in this example to average across items to arrive at summary dimension scores. Thus, any item-related variance will interfere with the intent to average across items and will inescapably contribute to score *un*reliability.

Out of the 25 studies that applied G coefficients in our review (see Table 9), only 16 studies presented at least one direct score representation. While illustrative G or Phi coefficients can potentially be helpful for demonstrating theoretical points, we suggest that direct representations of score reliability may be sought by researchers for many of the measures that are evaluated with G theory (e.g., to inform meta-analyses). There may be sound reasons why direct coefficients do not take precedence in a study where the researcher wishes to develop hypothetical arguments. Nonetheless, it would be helpful to make direct coefficients available when publishing G and D studies and, importantly, to clearly differentiate illustrative from direct coefficients (see Guideline 3A, Table 10).

To provide guidance on effect classification, if it is hypothetically feasible to aggregate across levels of facet y (e.g., aggregating across multiple raters or rating items) to arrive at intended construct scores, then y should normally be classified as a contribution to *un*reliable variance. For example, if it is feasible to aggregate ratings across different raters in

a 360-degree procedure, then any rater-related effects necessarily contribute to *un*reliability in a direct G coefficient. This does not imply that a researcher will wish to adjust for aggregation across levels of y when presenting their results. For example, it may be that outcomes are desired for a *given* (i.e., one) rater. However, the point is that even in this case, aggregation across raters is a feasible prospect. In addition to excluding from universe score variance effects implicated in aggregation, some effects are inherently inadmissible as contributions to universe score variance. This is because they represent measurement "scaffolding" unrelated to constructs of substantive interest (e.g., idiosyncratic rater or incidental item effects).

Reliability Estimates for Raw Observations. As a further point relating to illustrative reliability estimates, where it was possible to determine we found that 11 out of the 25 studies that had applied G theory reliability coefficients estimated them for raw observations (e.g., disaggregated ratings, item responses, etc.). Reliability estimates in G theory usually apply to summary scores rather than individual observations (Webb et al., 2006). Unless researchers have an illustrative reason to estimate reliability as it pertains to individual observations, a more cogent application involves presenting G or Phi coefficients for summary scores only (see Guideline 3B, Table 10).

Providing Detail Sufficient for Replicability. Unlike many other approaches to reliability estimation, G studies and their associated G and Phi coefficients are tailored to the specifics of a measurement design. Given the bespoke nature of G studies, it is important that researchers provide a level of detail sufficient to facilitate an understanding of what was done in a study and how the study can be reproduced (Köhler & Cortina, 2023). A key factor relevant to this issue in the context of G theory is that formulas for aggregation for G and Phi coefficients should be explicitly available in published works.

Under half of the studies in our review applied G study aggregation formulas (11/28). This is unfortunate because a major advantage of G theory is that the estimates it produces can be adjusted for aggregation effects. Of those studies where aggregation formulas were applied, most reported their formulas (10/11). The situation was however less encouraging for the application of G theory reliability coefficients. Twenty-five studies reported G coefficients, of which 20 applied aggregation formulas to those coefficients. Yet, G coefficient formulas were provided for only 16 studies. Without formulas, it is difficult or impossible to follow the process that was used to arrive at the estimates that are presented. Our recommendation is to make available formulas for (a) G study aggregation and (b) G or Phi coefficient aggregation. Doing so will enable readers to follow the logic adopted in a research paper relating to these key features of G theory and will assist in facilitating open-science replicability practices (see Guidelines 3C and 3D, Table 10).

Software. Another issue relevant to reproducibility concerns the software that was applied and the estimation basis for G studies and G or Phi coefficients. Regarding software, Table 8 shows that 11/28 studies did not specify the software used to conduct their analyses. Moreover, the estimator that was applied was not specified in 7/28 studies. Three studies mentioned VARCOMP, which refers to a group of algorithms available in the proprietary IBM SPSS and SAS packages. However, the VARCOMP term does not reveal the estimator that was used in the analysis. For example, in IBM SPSS (version 31), VARCOMP could refer to either MINQUE, maximum likelihood, ANOVA, or REML estimation (IBM Corp, 2025). Similarly, one study referenced the R program. However, this does not clarify details about the package that was used or its version. To facilitate conditions for reproducibility, we recommend that researchers declare (a) the software that was applied and its version, including details about R packages if relevant, as well as (b) the specific estimator that was

applied (see Guidelines 3E and 3F, Table 10). We provide further details about software relevant to G theory and the estimation basis for each package type in Table 11.

Additional Considerations

We found four studies in which Phi coefficients were reported in addition to G coefficients. As mentioned above, Phi coefficients are used when cut-scores or specific criteria are applied to pass an evaluation. In the papers where Phi coefficients were applied in our review, there was no clear indication that a set criterion for passing was relevant to the evaluation under scrutiny. We encourage researchers to either provide such context when applying Phi coefficients or, if the interest is concerned with individual differences, to report G coefficients only (see Guideline 3G, Table 10).

In a large multi-study application in our review, we found that unadjusted variance estimates were not reported or made available. This raises challenges for replication because the original variances are required if the aim is to verify any adjustments that have subsequently been applied to them. For this reason, we recommend that unadjusted variance estimates be made available to a precision of at least four decimal places, either in a publication or as supplemental material (see Guideline 3H, Table 10). This high level of precision is typically required for variance components analyses because variance estimates from these analyses are regularly small (Shavelson & Webb, 1991).

As a final consideration, in four of the studies we reviewed, we found errors in G coefficient equations that, of reassurance, were minor to the extent that their correction would not change the overall outcomes reported in their original papers. To help address this issue for future research, we provide additional guidance in our supplemental material for calculating G and Phi coefficients (see file “G Theory Coefficients.pdf”).

Summary and Conclusions

G theory is a complex paradigm, and no single treatment will do it justice. However, it is our hope that our tutorial-based examples, systematic review, and guidelines will help to demystify what G theory is and where and how it can be usefully applied. Our coverage of univariate G theory is relevant to every G theory application we were able to identify in the organizational context. This focus is intended to provide a basis for better informed future applications of the univariate approach. It moreover provides a platform for other applications such as multivariate G theory given that the multivariate framework is a logical extension of the univariate approach.

To provide a future oriented view, a chief advantage of G theory over other reliability perspectives is in its capacity to handle measurement designs of practically any level of complexity whilst retaining a coherent, consistent analytical foundation. With increasing complexity in organizational measurement, a flexible analytic paradigm such as G theory is needed more than ever before. The rise of artificial intelligence is a case in point where large language models can potentially evaluate data using more complex measurement procedures than those encountered previously. G theory presents a methodological opportunity well suited to this environment that may lead to its transition from relative obscurity into the mainstream.

References

*Indicates work included in the systematic review.

American Psychological Association. (2025). *Quantitative Research Design (JARS–Quant)*. https://apastyle.apa.org/jars/quantitative

*Arthur, W., Jr., Woehr, D. J., & Maldegan, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical reexamination of the assessment center construct-related validity paradox. *Journal of Management*, *26*(4), 813–835. https://doi.org/10.1016/S0149-2063(00)00057-X

Austin, P. C. (2010). Estimating multilevel logistic regression models when the number of clusters is low: a comparison of different statistical software procedures. *International Journal of Biostatistics*, *6*(1), 1–18. https://doi.org/10.2202/1557-4679.1195

*Baranowski, L. E., & Anderson, L. E. (2005). Examining rating source variation in work behavior to KSA linkages. *Personnel Psychology*, *58*(4), 1041–1054. https://doi.org/10.1111/j.1744-6570.2005.00234.x

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

*Blair, C. A., Gorman, C. A., Helland, K., & Delise, L. (2014). The smart leader: Examining the relationship between intelligence and leader development behavior. *Leadership and Organization Development Journal*, *35*(3), 241–258. https://doi.org/10.1108/LODJ-06-2012-0078

*Bowler, M. C., & Woehr, D. J. (2009). Assessment center construct-related validity: Stepping beyond the MTMM matrix. *Journal of Vocational Behavior*, *75*(2), 173–182. https://doi.org/10.1016/j.jvb.2009.03.008

*Breil, S. M., Forthmann, B., & Back, M. D. (2022). Measuring distinct social skills via multiple speed assessments: A behavior-focused personnel selection approach. *European Journal of Psychological Assessment*, *38*(3), 224–236. https://doi.org/10.1027/1015-5759/a000657

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x

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

Brennan, R. L. (2001b). *Manual for urGENOVA*.

Bürkner, P. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

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

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

*Dartt, A., Rosecrance, J., Gerr, F., Chen, P., Anton, D., & Merlino, L. (2009). Reliability of assessing upper limb postures among workers performing manufacturing tasks. *Applied Ergonomics*, *40*(3), 371–378. https://doi.org/10.1016/j.apergo.2008.11.008

*de Jong, A., Wetzels, M., & de Ruyter, K. (2008). Linking employee perceptions of collective efficacy in self-managing service teams with customer-perceived service quality: A psychometric assessment. *International Journal of Service Industry Management*, *19*(3), 353–378. https://doi.org/10.1108/09564230810875011

DeShon, R. P. (2002). Generalizability theory. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations* (pp. 189–220). Jossey-Bass.

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, *3*(4), 484–501. https://doi.org/10.1177/2515245920951747

Fyans, L. J., Jr. (1977). *A new multiple level approach to cross-cultural psychological research* [Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign].

*Greguras, G. J., Robie, C., Schleicher, D. J., & Goff, M., III. (2003). A field study of the effects of rating purpose on the quality of multisource ratings. *Personnel Psychology*, *56*(1), 1–21. https://doi.org/10.1111/j.1744-6570.2003.tb00141.x

*Herde, C. N., & Lievens, F. (2023). Multiple, speeded assessments under scrutiny: Underlying theory, design considerations, reliability, and validity. *Journal of Applied Psychology*, *108*(3), 351–373. https://doi.org/10.1037/apl0000603

*Highhouse, S., Broadfoot, A., Yugo, J. E., & Devendorf, S. A. (2009). Examining corporate reputation judgments with generalizability theory. *Journal of Applied Psychology*, *94*(3), 782–789. https://doi.org/10.1037/a0013934

Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Thomson Wadsworth.

IBM Corp. (2025). *IBM SPSS Statistics for Windows*. In (Version 31) [Computer software]. IBM Corp.

*Ingold, P. V., Heimann, A. L., Waller, B., Breil, S. M., & Sackett, P. R. (2025). What do assessment center ratings reflect? Consistency and heterogeneity in variance composition across multiple samples. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0001318

*Jackson, D. J. R., LoPilato, A. C., Hughes, D., Guenole, N., & Shalfrooshan, A. (2017). The internal structure of situational judgement tests reflects candidate main effects: Not dimensions or situations. *Journal of Occupational and Organizational Psychology*, *90*(1), 1–27. https://doi.org/10.1111/joop.12151

Jackson, D. J. R., Michaelides, G., Dewberry, C., & Englert, P. (2022). Clarifying the scope of generalizability theory for multifaceted assessment. *New Zealand Journal of Psychology*, *51*(2), 53–64.

*Jackson, D. J. R., Michaelides, G., Dewberry, C., Jones, A., Toms, S., Schwencke, B., & Yang, W.-N. (2022). Uncertainty about rater variance and small dimension effects impact reliability in supervisor ratings. *Human Performance*, *35*(3-4), 278–301. https://doi.org/10.1080/08959285.2022.2111433

*Jackson, D. J. R., Michaelides, G., Dewberry, C., Nelson, J., & Stephens, C. (2022). Reliability in assessment centres depends on general and exercise performance, but not on dimensions. *Journal of Occupational and Organizational Psychology*, *95*(4), 739–757. https://doi.org/10.1111/joop.12398

*Jackson, D. J. R., Michaelides, G., Dewberry, C., Schwencke, B., & Toms, S. (2020). The implications of unconfounding multisource performance ratings. *Journal of Applied Psychology*, *105*(3), 312–329. https://doi.org/10.1037/apl0000434

*Jackson, D. J. R., Michaelides, M., Dewberry, C., & Kim, Y. (2016). Everything that you have ever been told about assessment center ratings is confounded. *Journal of Applied Psychology*, *101*(7), 976–994. https://doi.org/10.1037/apl0000102

*Jackson, D. J. R., Stillman, J. A., & Atkins, S. G. (2005). Rating tasks versus dimensions in assessment centers: A psychometric comparison. *Human Performance*, *18*(3), 213–241. https://doi.org/10.1207/s15327043hup1803_2

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

Köhler, T., & Cortina, J. M. (2023). Constructive replication, reproducibility, and generalizability: Getting theory testing for JOMSR right. *Journal of Management Scientific Reports*, *1*(2), 75–93. https://doi.org/10.1177/27550311231176016

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Academic Press.

Kruschke, J. K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behavior*, *5*(10), 1282–1291. https://doi.org/10.1038/s41562-021-01177-7

*Lee, S., Lee, J., & Youn, C.-Y. (2005). A variation of CAT for measuring creativity in business products. *Korean Journal of Thinking & Problem Solving*, *15*(2), 143–153.

*Li, G., Pan, Y., & Wang, W. (2021). Using Generalizability theory and Many-facet Rasch Model to evaluate in-basket tests for managerial positions. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.660553

*Lievens, F. (2001a). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*, *86*(2), 255–264. https://doi.org/10.1037/0021-9010.86.2.255

*Lievens, F. (2001b). Assessors and use of assessment centre dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior*, *22*(3), 203–221. https://doi.org/10.1002/job.65

*Lievens, F., & Sanchez, J. I. (2007). Can training improve the quality of inferences made by raters in competency modeling? A quasi-experiment. *Journal of Applied Psychology*, *92*(3), 812–819. https://doi.org/10.1037/0021-9010.92.3.812

*Lievens, F., Sanchez, J. I., Bartram, D., & Brown, A. (2010). Lack of consensus among competency ratings of the same occupation: Noise or substance? *Journal of Applied Psychology*, *95*(3), 562–571. https://doi.org/10.1037/a0018035

*Lievens, F., Sanchez, J. I., & de Corte, W. (2004). Easing the inferential leap in competency modeling: The effects of task-related information and subject matter expertise. *Personnel Psychology*, *57*(4), 881–904. https://doi.org/10.1111/j.1744-6570.2004.00009.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*(2), 692–717. https://doi.org/10.1177/0149206314554215

Marcoulides, G. A. (1990). An alternative method for estimating variance components in generalizability theory. *Psychological Reports*, *66*(2), 379–386. https://doi.org/10.2466/PR0.66.2.379-386

Marcoulides, G. A. (1998). Applied generalizability theory models. In G. A. Marcoulides (Ed.), *Modern methods for business research.* (pp. 1–21). Lawrence Erlbaum Associates.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Novick, M. R. (1976). Bayesian methods in educational testing: A third survey. In D. M. N. De Gruijter & L. J. T. van der Kamp (Eds.), *Advances in psychological and educational testing* (pp. 17–32). Wiley.

*O'Neill, T. A., McLarnon, M. J. W., & Carswell, J. J. (2015). Variance components of job performance ratings. *Human Performance*, *28*(1), 66–91. https://doi.org/10.1080/08959285.2014.974756

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., . . . Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*, 1–9. https://doi.org/10.1136/bmj.n71

*Park, K. W., Arvey, R. D., & Tong, Y. K. (2011). The generalizability of leadership across activity domains and time periods. *The Leadership Quarterly*, *22*(1), 223–237. https://doi.org/10.1016/j.leaqua.2010.12.017

*Putka, D. J., & Hoffman, B. J. (2013). Clarifying the contribution of assessee-, dimension-, exercise-, and assessor-related effects to reliable and unreliable variance in assessment center ratings. *Journal of Applied Psychology*, *98*(1), 114–133. https://doi.org/10.1037/a0030887

Putka, D. J., & Hoffman, B. J. (2014). "The" reliability of job performance ratings equals 0.52. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 247–275). Taylor & Francis.

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: implications for estimating interrater reliability. *Journal of Applied Psychology*, *93*(5), 959–981. https://doi.org/10.1037/0021-9010.93.5.959

Raykov, T., & Marcoulides, G. A. (2006). Estimation of generalizability coefficients via a structural equation modeling approach to scale reliability evaluation. *International Journal of Testing*, *6*(1), 81–95. https://doi.org/10.1207/s15327574ijt0601_5

Raykov, T., & Marcoulides, G. A. (2011). *An introduction to psychometric theory*. Routledge/Taylor & Francis Publishers.

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement*, *79*(1), 200–210. https://doi.org/10.1177/0013164417725127

*Robie, C., Born, M. P., & Schmit, M. J. (2001). Personal and situational determinants of personality responses: A partial reanalysis and reinterpretation of the Schmit et al. (1995) data. *Journal of Business and Psychology*, *16*(1), 101–117. https://doi.org/10.1023/A:1007843906550

Rudin, W. (1987). *Real and complex analysis* (3rd ed.). McGraw-Hill.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology*, *53*(4), 901–912. https://doi.org/10.1111/j.1744-6570.2000.tb02422.x

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. Wiley.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology*, *34*, 133–166. https://doi.org/10.1111/j.2044-8317.1981.tb00625.x

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

Shavelson, R. J., & Webb, N. M. (2005). Generalizability theory. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Complementary methods for research in education* (3rd ed., pp. 599–612). AERA.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428. https://doi.org/10.1037/0033-2909.86.2.420

ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2024). Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. *Psychological Methods*, *29*(5), 967–979. https://doi.org/10.1037/met0000516

Vispoel, W. P., Lee, H., Chen, T., & Hong, H. (2024). Analyzing and comparing univariate, multivariate, and bifactor generalizability theory designs for hierarchically structured

personality traits. *Journal of Personality Assessment*, *106*(3), 285–300. https://doi.org/10.1080/00223891.2023.2268193

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

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 81–124). Elsevier B. V. https://doi.org/10.1016/S0169-7161(06)26004-8

Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An examination of G-Theory methods for modeling multitrait–multimethod data: Clarifying links to construct validity and confirmatory factor analysis. *Organizational Research Methods*, *15*(1), 134–161. https://doi.org/10.1177/1094428111408616

Table 1

Key Terms for Generalizability Theory

Term	Definition
Absolute decisions	Where cut-scores or pass/fail criteria are used in scoring
Aggregation	Where observations (e.g., ratings) are summarized (e.g., averaged)
Balanced versus unbalanced designs	A balanced design is where equal numbers of levels are nested in levels of another effect (e.g., each of 5 dimensions is measured with 3 items). An unbalanced design is where nested levels are unequal.
Crossed effects	Where all levels of one effect occur across all levels of another effect (e.g., all people responded to all items)
Decision (D) study	Where Generalizability (G) study results are used to estimate G and/or Phi coefficients
Effect	A related group of measurement subcomponents (or *levels*, e.g., *n* items) often modeled with a variance component
Error variance	Variance that contributes to measurement unreliability
Facet	A source of variability that may impact measurement unreliability (e.g., idiosyncratic rater variance)
Fixed effect	Where the researcher is interested in comparing specific levels in an effect (e.g., I want to know if Claude scores higher than ChatGPT on a test)
Focus (object) of measurement	The source of variability that represents what or whomever is being evaluated (e.g., people, video recorded performances)
G coefficient	A G theory reliability coefficient for relative decisions
G study	A decomposition of sources of reliable, unreliable, or irrelevant variance
Interaction	The multiplicative combination of > 1 main effects (e.g., Persons × Raters)
Levels	The subcomponents of an effect (e.g., *n* individual items that comprise the items effect or all the *n* people who comprise the participant effect)
Main effect	An estimate for a single variance component (e.g., persons)
Multifaceted measurement design	A measurement procedure that includes multiple different facets (e.g., items, raters, occasions)
Nested effects	Where unique sets of levels for one effect are grouped into levels on another effect (e.g., subgroups of *n* unique items are grouped into Five Factor Model traits)
Phi coefficient	A G theory reliability coefficient for absolute decisions
Random effect	Where it is assumed that all levels of an effect have been sampled from a larger universe of such levels or are at least exchangeable with other, similar levels (e.g., my raters could reasonably be exchanged with others)
Relative decisions	Where the measurement focus is on individual differences (e.g., normed comparisons, comparing participants, a focus on participant variability)
Summary score	An average or summation of observations (e.g., averaged item responses)
Universe	A potentially infinite number of levels, often reserved for describing facets[a] (e.g., a potentially infinite number of items deemed to measure well-being)
Universe of admissible observations	Observations that a researcher is willing to treat as interchangeable (e.g., my well-being items could reasonably be exchanged with other items from a hypothetical universe of admissible well-being items)
Universe of generalization	The set of relevant measurement conditions to which the researcher seeks to generalize based on their measurement procedure (e.g., occasions [repeated measures] may be relevant to a measure used for employee development but not for employee selection)
Universe score variance	The G theory analogue of true score or reliable variance, represented by the hypothetical average score across a potentially infinite number of levels from a relevant set of facets (e.g., the average across a hypothetically infinite number of ability test items and occasions of measurement)

Note. [a]Strictly speaking, the term "population" is reserved for the *focus (object) of measurement*.

Table 2

Comparisons and Data Considerations for Generalizability Theory Estimators

Estimation for an example $p \times i$ design					
Random effect	Traditional ANOVA computations		Alternative Estimator Comparison		
	Variance component	Mean Squares (MS)	Traditional ANOVA	REML	Bayesian
Residual	$\sigma^2_{pi,e} = MS_{pi,e}$	$MS_{pi,e} = SS_{pi,e}/df_{pi,e}$	Estimated via subtraction from total sums of squares	Estimated as a residual variance model parameter	Estimated as a full posterior distribution for residual variance
Item main effect	$\sigma^2_i = (MS_i - \sigma^2_{pi,e})/n_p$	$MS_i = SS_i/df_i$	Derived from mean squares based on item variance	Estimated as an item variance model parameter	Estimated as a full posterior distribution for item variance
Person main effect	$\sigma^2_p = (MS_p - \sigma^2_{pi,e})/n_i$	$MS_p = SS_p/df_p$	Derived from mean squares based on person variance	Estimated as a person variance model parameter	Estimated as a full posterior distribution for person variance
General data considerations			Traditional ANOVA	REML	Bayesian
Balanced data			Required unless using Henderson's Method I	Not required	Not required
Full crossing			Required if relevant	Not required	Not required
Complete data			Required	Not required	Not required
Estimation of uncertainty			Confidence intervals for non-negative effects	Confidence intervals for non-fenced effects	Credible intervals for all parameters
Normality assumptions			Not assumed	Assumed	Determined by model likelihood
Negative or zero-fenced variance			Negative variances possible	Zero-fencing possible	Avoided with appropriate prior specification

Note. This design is the generalizability theory equivalent of coefficient alpha. REML = restricted maximum likelihood. σ^2 = variance component, p = person, i = item, e = residual error, *MS* = mean square, *df* = degrees of freedom, estimated as follows: residual = $(n_p - 1)(n_i - 1)$, item main effect = $n_i - 1$, and person main effect = $n_p - 1$.

Table 3

Example 1: Generalizability and Decision Studies for p × i Design

Effect	n_{total}	95LB	VC	95UB	Tot%	Overall (O) Scores				
						O-*f*	O-rel	O-rel%	O-abs	O-abs%
p	100	0.6303	0.8540	1.1713	64	p	0.8540	87	0.8540	84
i	3	0.0148	0.1198	1.4428	9	i/n_i			0.0399	4
pi,*e*	300	0.3014	0.3660	0.4451	27	pi,*e*/n_i	0.1220	13	0.1220	12
G (O Score)		0.8265	0.8753	0.9112						
Phi (O Score)		0.5622	0.8354	0.8934						
Coefficient alpha		0.8263	0.8755	0.9125						

Note. Relates to supplemental p × i design code. b95LB–95UB = 95% credible intervals for generalizability theory estimates and confidence intervals for comparator estimates, VC = variance component point estimate, Tot% = proportion of total variance explained in ratings. *f* = aggregation formula for overall (O) summary scores, O-rel = effects adjusted for aggregation for relative decisions (i.e., individual differences), O-abs = effects adjusted for aggregation for absolute decisions (i.e., use of cut scores or absolute criteria). p = participant followers, d = dimensions, i = rating items, *e* = residual error, : = a level of nesting. G = generalizability coefficient, Phi = Phi coefficient. Coefficient alpha is presented as a comparator for the G (O Score) estimate. Formulas for G and Phi coefficients – G (O Score) from O-rel column = p / p + remaining effects; Phi (O Score) from O-abs column = p / p + remaining effects.

Table 4

Example 2: Generalizability and Decision Studies for Unbalanced p × (i:d) Design

Effect	n_{total}	95LB	VC	95UB	Tot%	Dimension (D) Scores			Overall (O) Scores				
						D-*f*	D-rel	D-rel%	O-*f*	O-rel	O-rel%	O-abs	O-abs%
p	100.00	0.9772	1.2883	1.7677	44	p	1.2883	74	p	1.2883	93	1.2883	84
d	5.00	0.1911	0.7419	3.1649	26				d/n_d			0.1484	10
i:d	24.00	0.0414	0.1031	0.3288	4				i:d/$n_{i:d}n_d$			0.0047	<1
pd	500.00	0.3037	0.3674	0.4481	13	pd	0.3674	21	pd/n_d	0.0735	5	0.0735	5
pi:d,*e*	2,400.00	0.3626	0.3966	0.4332	14	pi:d,*e*/$n_{i:d}$	0.0907	5	pi:d,*e*/$n_{i:d}n_d$	0.0181	1	0.0181	1
$n_{i:d}$[a]	4.38												
G (D Score)		0.9353	0.9483	0.9600									
G (O Score)		0.9102	0.9338	0.9518									
Phi (O Score)		0.6316	0.8393	0.9112									
Omega		0.9722	0.9835	0.9904									

Note. Relates to supplemental p × (i:d) design code. 95LB–95UB = 95% credible intervals for generalizability theory estimates and confidence intervals for comparator estimates, VC = variance component point estimate, Tot% = proportion of total variance explained in ratings. *f* = aggregation formula for dimension (D) or overall (O) summary scores, D-rel and O-rel = effects adjusted for aggregation for relative decisions (i.e., individual differences), O-abs = effects adjusted for aggregation for absolute decisions (i.e., use of cut scores or absolute criteria). To preserve clarity, absolute decision information is only presented for overall scores. D-rel% and O-abs% = percentage of variance explained post aggregation adjustment for relative and absolute decisions respectively. p = participant followers, d = dimensions, i = rating items, *e* = residual error, : = a level of nesting. [a]Represents harmonic mean $n_{i:d}$, with item *n* per dimension = 3, 4, 7, 4, 6. G = generalizability coefficient, Phi = Phi coefficient. Omega = McDonald's composite omega based on a five-dimension confirmatory factor analysis. Omega is presented as a comparator for G (D Score). Formulas for G and Phi coefficients – G (D Score) from D-rel column = p + pd / p + pd + remaining effect; G (O Score) from O-rel column = p / p + remaining effects; Phi (O Score) from O-abs column = p / p + remaining effects.

Table 5

Example 3: Generalizability and Decision Study for (p:l) × i Design

Effect	n_{total}	95LB	VC	95UB	Tot%	Overall scores level 1: focus of measurement = followers (p:l)					Overall scores level 2: focus of measurement = leaders (l)				
						O1-*f*	O1-rel	O1-rel%	O1-abs	O1-abs%	O2-*f*	O2-rel	O2-rel%	O2-abs	O2-abs%
l	100	0.5979	0.7895	1.0787	39						l	0.7895	96	0.7895	89
p:l	3,000	0.4696	0.5036	0.5399	25	p:l	0.5036	76	0.5036	69	p:l/$n_{p:l}$	0.0168	2	0.0168	2
i	3	0.0321	0.1972	1.8299	10	i/n_i			0.0657	9	i/n_i			0.0657	7
li	300	0.0335	0.0437	0.0575	2						li/n_i	0.0146	2	0.0146	2
pi:l,e	9,000	0.4741	0.4914	0.5099	24	pi:l,e/n_i	0.1580	24	0.1638	22	pi:l,e/$n_{p:l}n_i$	0.0055	1	0.0055	1
$n_{p:l}$	30														
G (O1 Score)		0.7389	0.7544	0.7699											
G (O2 Score)		0.9405	0.9555	0.9672											
Phi (O1 Score)		0.3921	0.6857	0.7469											
Phi (O2 Score)		0.5522	0.8825	0.9479											
ML omega(O1)		0.7381	0.7553	0.7720											
ML omega(O2)		0.9747	0.9823	0.9854											

Note. Relates to supplemental (p:l) × i design code. Results are presented for two levels, including level 1 (participant followers nested in leaders, p:l) and level 2 (leaders, l). 95LB–95UB = 95% credible intervals for generalizability theory estimates and confidence intervals for comparator estimates, VC = variance component point estimate, Tot% = proportion of total variance explained in ratings. *f* = aggregation formula for dimension (D) or overall (O) summary scores, D-rel and O-rel = effects adjusted for aggregation for relative decisions (i.e., individual differences), O-abs = effects adjusted for aggregation for absolute decisions (i.e., use of cut scores or absolute criteria). To preserve clarity, absolute decision information is only presented for overall scores. D-rel% and O-abs% = percentage of variance explained post aggregation adjustment for relative and absolute decisions respectively. p = participant followers, d = dimensions, i = rating items, *e* = residual error, : = a level of nesting. $n_{p:l}$ denotes number of p nested in each level of l. G = generalizability coefficient, Phi = Phi coefficient. ML Omega = McDonald's multilevel omega based on 3 items loading on an overall factor for level 1 (p:l) and level 2 (l) via confirmatory factor analysis. ML Omega O1 and O2 are presented as comparators for G (O1 Score) and G (O2 Score) respectively. Formulas for G and Phi coefficients – G (O1 Score) from O1-rel column = p:l / p:l + remaining effects; G (O2 Score) from O2-rel column = l / l + remaining effects; Phi (O1 Score) from O1-abs column = p:l / p:l + remaining effects; Phi (O2 Score) from O2-abs column = l / l + remaining effects.

Table 6

Example 4: Generalizability and Decision Study for Balanced (p:l) × (i:d) Design for Dimension Scores

Effect	n_{total}	95LB	VC	95UB	Tot%	Dimension scores level 1: focus of measurement = followers (p:l)			Dimension scores level 2: focus of measurement = leaders (l)		
						D1-*f*	D1-rel	D1-rel%	D2-*f*	D2-rel	D2-rel%
l	100	0.5549	0.7267	0.9794	27				l	0.7267	82
p:l	3,000	0.3337	0.3563	0.3808	13	p:l	0.3563	40	p:l/$n_{p:l}$	0.0119	1
d	5	0.0652	0.4023	2.0068	15						
i:d	15	0.0761	0.1801	0.5260	7						
ld	500	0.0782	0.0968	0.1192	4				ld	0.0968	11
li:d	1,500	0.0816	0.0904	0.1002	3				li:d/$n_{i:d}$	0.0301	3
pd:l	15,000	0.3531	0.3665	0.3808	14	pd:l	0.3665	41	pd:l/$n_{p:l}$	0.0122	1
pi:ld, e	45,000	0.4823	0.4900	0.4982	18	pi:ld,e/$n_{i:d}$	0.1608	18	pi:ld,e/$n_{p:l}n_{i:d}$	0.0054	1
$n_{p:l}$	30										
$n_{i:d}$	3										
G (D1 Score)		0.8095	0.8157	0.8219							
G (D2 Score)		0.9157	0.9325	0.9478							
ML omega (D1)		0.9252	0.9294	0.9326							
ML omega (D2)		0.9892	0.9920	0.9942							

Note. Relates to supplemental (p:l) × (i:d) design code. Results are presented for two levels, including level 1 (participant followers nested in leaders, p:l) and level 2 (leaders, l). 95LB–95UB = 95% credible intervals for generalizability theory estimates and confidence intervals for comparator estimates, VC = variance component point estimate, Tot% = proportion of total variance explained in ratings. *f* = aggregation formula for dimension (D) summary scores, D-rel = effects adjusted for aggregation for relative decisions (i.e., individual differences). D-rel% = percentage of variance explained post aggregation adjustment for relative decisions. p = participant followers, d = dimensions, i = rating items, *e* = residual error, : = a level of nesting. $n_{p:l}$ = number of p nested in each level of l and $n_{i:d}$ = number of i nested in each level of d. G = generalizability coefficient. ML Omega = McDonald's multilevel omega based on five-factor solutions for level 1 (p:l) and level 2 (l) via confirmatory factor analysis. ML Omega D1 and D2 are presented as comparators for G (D1 Score) and G (D2 Score) respectively. Formulas for G coefficients – G (D1 Score) from D1-rel column = p:l + pd:l / p:l + pd:l + remaining effect; G (D2 Score) from D2-rel column = l + ld / l + ld + remaining effects.

Table 7

Example 4: Generalizability and Decision Study for Balanced (p:l) × (i:d) Design for Overall Scores

Effect	n_{total}	95LB	VC	95UB	Tot%	Overall scores level 1: focus of measurement = followers (p:l)					Overall scores level 2: focus of measurement = leaders (l)				
						O1-*f*	O1-rel	O1-rel%	O1-abs	O1-abs%	O2-*f*	O2-rel	O2-rel%	O2-abs	O2-abs%
l	100	0.5549	0.7267	0.9794	27						l	0.7267	95	0.7267	85
p:l	3,000	0.3337	0.3563	0.3808	13	p:l	0.3563	77	0.3563	64	p:l/$n_{p:l}$	0.0119	2	0.0119	1
d	5	0.0652	0.4023	2.0068	15	d/n_d			0.0805	15	d/n_d			0.0805	9
i:d	15	0.0761	0.1801	0.5260	7	i:d/$n_{i:d}n_d$			0.0120	2	i:d/$n_{i:d}n_d$			0.0120	1
ld	500	0.0782	0.0968	0.1192	4						ld/n_d	0.0194	3	0.0194	2
li:d	1,500	0.0816	0.0904	0.1002	3						li:d/$n_{i:d}n_d$	0.0060	1	0.0060	1
pd:l	15,000	0.3531	0.3665	0.3808	14	pd:l/n_d	0.0733	16	0.0733	13	pd:l/$n_{p:l}n_d$	0.0024	<1	0.0024	<1
pi:ld, e	45,000	0.4823	0.4900	0.4982	18	pi:ld,e/$n_{i:d}n_d$	0.0327	7	0.0327	6	pi:ld,e/$n_{p:l}n_{i:d}n_d$	0.0011	<1	0.0011	<1
$n_{p:l}$	30														
$n_{i:d}$	3														
G (O1 Score)		0.7575	0.7708	0.7835											
G (O2 Score)		0.9302	0.9469	0.9608											
Phi (O1 Score)		0.4042	0.6407	0.7287											
Phi (O2 Score)		0.6047	0.8433	0.9189											
ML ICC2 (O1)		0.9176	0.9217	0.9259											
ML ICC2 (O2)		0.9720	0.9796	0.9845											

Note. Results are presented for two levels, including level 1 (participant followers nested in leaders, p:l) and level 2 (leaders, l). 95LB–95UB = 95% credible intervals for generalizability theory estimates and confidence intervals for comparator estimates, VC = variance component point estimate, Tot% = proportion of total variance explained in ratings. *f* = aggregation formula for overall (O) summary scores, O-rel = effects adjusted for aggregation for relative decisions (i.e., individual differences), O-abs = effects adjusted for aggregation for absolute decisions (i.e., use of cut scores or absolute criteria). O-rel% = percentage of variance explained post aggregation adjustment for relative decisions. p = participant followers, d = dimensions, i = rating items, *e* = residual error, : = a level of nesting. $n_{p:l}$ = number of p nested in each level of l and $n_{i:d}$ = number of i nested in each level of d. G = generalizability coefficient. ML ICC2 = the multilevel form of intraclass correlation 2 based on Shrout and Fleiss (1979), including level 1 for followers, p:l, adjusted for n_i and level 2 for leaders, l, adjusted for $n_{p:d}$ and n_i. ML ICC2 O1 and O2 are presented as comparators for G (O1 Score) and G (O2 Score) respectively. Formulas for G coefficients – G (O1 Score) from O1-rel column = p:l / p:l + remaining effect; G (O2 Score) from O2-rel column = l / l + remaining effects; Phi (O1 Score) from O1-abs column = p:l / p:l + remaining effect; Phi (O2 Score) from O2-abs column = l / l + remaining effects.

Table 8

Summary Frequencies from Systematic Review of Generalizability Theory Applications in Organizational Research

Category	Frequency (total $k = 28$)
Journal publication	
Single-entry others[a]	11
Journal of Applied Psychology	9
Personnel Psychology	3
Human Performance	3
Journal of Occupational and Organizational Psychology	2
Estimator	
Frequentist	22
Bayesian	6
Software	
Unspecified	11
Stan/Rstan/brms	6
GENOVA	3
urGENOVA	2
SAS	2
lme4 (R package)	2
SPSS	1
R (package unspecified)	1
Estimators used to generate random effects	
Unspecified	7
Analogous-ANOVA	7
Markov Chain Monte Carlo Simulation	6
Restricted Maximum Likelihood	4
VARCOMP (unspecified)	3
MIVQUE	1
Generalizability theory keywords	
None	16
Generalizability theory	6
Reliability	3
Variance partitioning or components	2
Bayesian generalizability theory	1

Note. Analogous-ANOVA = Henderson's Method I, MIVQUE = minimum variance quadratic unbiased estimation, which is not to be confused with but is often practically equivalent to MINQUE (minimum norm quadratic unbiased estimation). MIVQUE and MINQUE do not produce standard error estimates and may generate negative variances. [a]Single cases of generalizability theory applications were published in the *Journal of Management, Leadership and Organization Development Journal, Journal of Vocational Behavior, European Journal of Psychological Assessment, Applied Ergonomics, International Journal of Service Industry Management, Korean Journal of Thinking and Problem Solving, Frontiers in Psychology, Journal of Organizational Behavior, Leadership Quarterly*, and the *Journal of Business and Psychology*. References included in our systematic review appear in our online supplement.

Table 9

Frequencies by Research Context for Generalizability Studies Included in Systematic Review

Research context	*k*	G Studies and Random Effects			G Coefficients					Measurement Issues	
		G study aggregation applied	G study aggregation formula present	G study CIs or Crls	G coef(s) applied	G coef aggregation applied	G coef formula present	≥ 1 G coef direct score[a]	G coef CIs or Crls	All *n* focus of measurement ≥ 30[b]	Evaluated Multiple constructs
ACs	10	5	5	5	9	8	7	6	1	8	10
Competencies	3	0	0	0	3	2	0	1	0	1	2
Performance ratings	2	1	1	0	2	1	1	1	0	2	2
360-degree ratings	2	1	1	0	2	1	2	2	1	2	2
Work simulations	2	0	0	1	2	1	1	2	0	2	1
Leadership	2	0	0	0	2	2	2	1	0	2	2
Job analysis	1	0	0	1	0	0	0	0	0	1	1
Corporate reputation	1	0	0	0	1	1	0	1	0	0	0
Personality	1	0	0	0	0	0	0	0	0	1	1
SJTs	1	1	1	0	1	1	1	0	0	1	1
Creativity	1	1	0	0	1	1	0	0	0	0	1
Service management	1	1	1	0	1	1	1	1	0	0	0
Body postures	1	1	1	0	1	1	1	1	0	0	0
Total *k*	28	11	10	7	25	20	16	16	2	20	23

Note. ACs = assessment centers, SJTs = situational judgment tests, G = generalizability, G coef = generalizability (reliability) coefficient. [a]By "direct score", we refer to estimates that are non-illustrative and approximate reliability for a specific score type. [b]Minimum required *n* (but not a guarantee) for approximating a data distribution under Central Limit Theorem (Howell, 2007).

Table 10

Guidelines for Conducting and Reporting Generalizability and Decision Studies

1.	Research design
A.	Ensure a sufficient *n* for key main effects (usually *n* = 30 is the bare minimum requirement), particularly for the focus (object) of measurement.
B.	Report *n* for each random effect.
C.	Apply estimators that can handle sparse data sets if relevant (e.g., Bayesian or REML procedures). Avoid deleting data to handle sparsity.
D.	Use univariate G theory for single-construct measures or relevant univariate interactions for summarizing average construct effects. Avoid applying separate univariate G studies for each construct.
E.	For multidimensional measures, apply univariate G theory to estimate reliability for any given dimension and apply multivariate G theory to estimate reliability for each separate dimension.
2.	Analytical procedures
A.	Consider applying Bayesian rather than Frequentist methods, given their advantages in the G theory context (e.g., they avoid negative variance estimates).
B.	To avoid Jensen's inequality when summarizing Bayesian results, compute squared quantities directly from posterior draws rather than squaring summary statistics (e.g., posterior medians or standard deviations).
C.	Report 95% credible intervals (for Bayesian analysis) or 95% confidence intervals (for Frequentist analysis) for random effects.
D.	When using Bayesian analysis, report credible intervals for G and Phi coefficients.
E.	If relevant (e.g., assessors and assesses partially overlap in a measurement design), apply the *q*-multiplier to relevant effects[a].
F.	When adjusting for aggregation for unbalanced univariate designs, apply and report the harmonic mean as an indication of central tendency for unbalanced nested effects.
G.	Consider applying D studies for reliability projections where like-with-like comparisons are of assistance (e.g., number of items differ between studies) or where reliability projections are helpful (e.g., what would happen if I averaged across a higher number of raters than I currently have in my procedure?).
3.	Presentation of results
A.	Distinguish between reliability estimates intended to develop illustrative points versus those intended to represent direct reliability estimates for scores. Consider prioritizing direct score reliability estimates.
B.	Report G and Phi coefficients for summary scores rather than for individual observations.
C.	Report aggregation formulas for random effects.
D.	Report aggregation formulas for G and Phi coefficients.
E.	Report software and software version used to analyze data.
F.	Report estimator type applied in data estimation.
G.	Apply G coefficients where interest is in individual differences and/or Phi coefficients where interest is in predetermined passing criteria and explain the context for either application.
H.	Provide original variance estimates for replicability to at least four decimal places.

Note. G = generalizability, D = decision, REML = restricted maximum likelihood. [a]Details about the *q*-multiplier can be found in Putka et al. (2008).

Table 11

Software Options for Generalizability Studies

Software	Availability	Inference	Estimator basis	Type	Response
facet	Github	Frequentist or Bayesian	ANOVA, REML, MCMC	Specialized	Any
GENOVA	University of Iowa	Frequentist	ANOVA	Specialized	Univariate
urGENOVA	University of Iowa	Frequentist	ANOVA	Specialized	Univariate
mGENOVA	University of Iowa	Frequentist	ANOVA	Specialized	Multivariate
G-String VI	Github	Frequentist	ANOVA (via urGENOVA)	Specialized	Univariate
G-String MV	Github	Frequentist	ANOVA	Specialized	Multivariate
GeneralizIT	Python package	Frequentist	ANOVA	Specialized	Univariate
lme4 – lmer()	R package	Frequentist	REML	Generic	Univariate
SAS – PROC VARCOMP	SAS package	Frequentist	MINQUE, ML, ANOVA, REML	Generic	Univariate
SPSS – VARCOMP	SPSS package	Frequentist	MINQUE, ML, ANOVA, REML	Generic	Univariate
brms	R package	Bayesian	MCMC	Generic	Any
rstan	R package	Bayesian	MCMC	Generic	Any
Stan	Stan package	Bayesian	MCMC	Generic	Any
WinBUGS	WinBUGS package	Bayesian	MCMC	Generic	Any
OpenBUGS	OpenBUGS package	Bayesian	MCMC	Generic	Any
MultiBUGS	MultiBUGS package	Bayesian	MCMC	Generic	Any

Note. ANOVA refers to analogous ANOVA procedures or specifically Henderson's Method I, REML = restricted maximum likelihood, MINQUE = minimum norm quadratic unbiased estimation, ML = maximum likelihood, MCMC = Markov Chain Monte Carlo simulation. All the above facilities require raw data. Not included above are the EduG specialized facility and the gtheory R package, both of which are currently unavailable. All solutions above handle balanced and unbalanced designs, apart from GENOVA, which handles balanced designs only. For example, if a measure has five constructs and five items nested in each construct, then the design is balanced. Conversely, if a measure has five constructs, but the item numbers differ per construct, then the design is unbalanced.

Appendix A

A Brief Primer on Bayesian Methods

We present Bayesian outcomes in our article because of their advantages in a G theory context (see Table 2) and because of the long deliberated alignment between Bayesian reasoning and G theory. Novick (1976) was one of the first to point out that "Generalizability Theory … is Bayesian in everything but a formal sense…" (p. 24). Fyans (1977) expanded on this notion by providing a general strategy for obtaining Bayesian G theory estimates for a variety of measurement designs (Shavelson & Webb, 1981).

Bayesian methods provide a principled approach towards updating a researcher's beliefs in light of evidence. This idea is formally stated as Posterior $\propto$ Likelihood $\times$ Prior, meaning that what you believe after observing data (the posterior) is proportional ($\propto$) to how well observed data fit your hypothesis (the likelihood), multiplied by what you believed before observing it (the prior). Bayesian analysis begins with a formal specification of prior beliefs about the parameters of a model. This is specified as a statistical distribution that describes these beliefs as probabilities (e.g., regression coefficients follow a normal distribution with a mean of 0 and a standard deviation of 1). Bayesian analysis is based on Markov Chain Monte Carlo simulations and updates prior beliefs with information drawn from observed data. This combination of priors and empirical updates results in a *posterior distribution* that represents results for all model parameters, which can be summarized with a mean or median, which serves as the analogue of a Frequentist point estimate. The lower and upper 95% percentiles of each parameter serve as a margin of error for the mean or median, referred to as a *credible interval*. This is roughly analogous to frequentist confidence intervals, but with a more intuitive interpretation reflecting the range of values containing the true parameter value with a specific probability given the observed data. The reader is directed to Kruschke (2014) for a more comprehensive treatment of Bayesian methods.

Appendix B

Generalizability and Phi Coefficient Equations for Examples 1 through 3

Example, score	Equation: Numerator = universe, denominator = universe + error	Key points to note
1, G (O Score)	$\frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pi,e}^2}{n_i}}$	Universe score = σ_p^2. Division by n_i for overall score aggregation across items. Relative decisions, so σ_i^2 is irrelevant.
1, Phi (O Score)	$\frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_i^2}{n_i} + \frac{\sigma_{pi,e}^2}{n_i}}$	Similar to the above equation, except absolute decisions mean that $\frac{\sigma_i^2}{n_i}$ contributes to error.
2, G (D Score)	$\frac{\sigma_p^2 + \sigma_{pd}^2}{\sigma_p^2 + \sigma_{pd}^2 + \frac{\sigma_{pi:d,e}^2}{n_{i:d}}}$	Dimension variance, σ_{pd}^2, contributes with σ_p^2 to universe. Division by $n_{i:d}$ reflects n items nested in each dimension.
2, G (O Score)	$\frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pd}^2}{n_d} + \frac{\sigma_{pi:d,e}^2}{n_{i:d}n_d}}$	Because of aggregation across dimensions, dimension effects, $\frac{\sigma_{pd}^2}{n_d}$, contribute to error.
2, Phi (O Score)	$\frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_d^2}{n_d} + \frac{\sigma_{i:d}^2}{n_{i:d}n_d} + \frac{\sigma_{pd}^2}{n_d} + \frac{\sigma_{pi:d,e}^2}{n_{i:d}n_d}}$	Absolute decisions, thus $\frac{\sigma_d^2}{n_d} + \frac{\sigma_{i:d}^2}{n_{i:d}n_d}$ contribute to error. Division reflects aggregation across all items and dimensions.
3, G (O1 Score)	$\frac{\sigma_{p:l}^2}{\sigma_{p:l}^2 + \frac{\sigma_{pi:l,e}^2}{n_i}}$	Follower focus, thus $\sigma_{p:l}^2$, = universe. Aggregation across items only.
3, G (O2 Score)	$\frac{\sigma_l^2}{\sigma_l^2 + \frac{\sigma_{p:l}^2}{n_{p:l}} + \frac{\sigma_{l:i}^2}{n_i} + \frac{\sigma_{pi:l,e}^2}{n_{p:l}n_i}}$	Leader focus, thus σ_l^2 = universe. Aggregation across followers (p) and items (i).
3, Phi (O1 Score)	$\frac{\sigma_{p:l}^2}{\sigma_{p:l}^2 + \frac{\sigma_i^2}{n_i} + \frac{\sigma_{pi:l,e}^2}{n_i}}$	Same as 3, G overall 1, except $\frac{\sigma_i^2}{n_i}$ contributes to error due to absolute decisions.
3, Phi (O2 Score)	$\frac{\sigma_l^2}{\sigma_l^2 + \frac{\sigma_{p:l}^2}{n_{p:l}} + \frac{\sigma_i^2}{n_i} + \frac{\sigma_{l:i}^2}{n_i} + \frac{\sigma_{pi:l,e}^2}{n_{p:l}n_i}}$	Same as 3, G overall 2, except $\frac{\sigma_i^2}{n_i}$ contributes to error due to absolute decisions.

Note. Universe = universe score variance, error = measurement error variance. Generalizability (G) coefficients are specified for relative decisions, and Phi coefficients are specified for absolute decisions. Universe score variance always appears in the numerator. Total variance relevant to the equation always appears in the denominator. Therefore, each G and Phi coefficient represents the proportion of universe score variance to total variance. Focus = focus (or object) of measurement. p = participants or followers, i = rating items, l = leaders, d = dimensions, *e* = residual error, : = one level of nesting.

Appendix C

Generalizability and Phi Coefficient Equations for Example 4

Example, score	Equation: Numerator = universe, denominator = universe + error	Key points to note
4, G (D1 Score)	$\frac{\sigma^2_{p:l} + \sigma^2_{pd:l}}{\sigma^2_{p:l} + \sigma^2_{pd:l} + \frac{\sigma^2_{pi:ld,e}}{n_{i:d}}}$	Follower focus with dimension scores, thus $\sigma^2_{p:l} + \sigma^2_{pd:l}$ = universe. Aggregation across items nested in each dimension.
4, G (D2 Score)	$\frac{\sigma^2_{l} + \sigma^2_{ld}}{\sigma^2_{l} + \sigma^2_{ld} + \frac{\sigma^2_{p:l}}{n_{p:l}} + \frac{\sigma^2_{li:d}}{n_{i:d}} + \frac{\sigma^2_{pd:l}}{n_{p:l}} + \frac{\sigma^2_{pi:ld,e}}{n_{p:l}n_{i:d}}}$	Leader focus with dimension scores, thus $\sigma^2_{l} + \sigma^2_{ld}$ = universe. Aggregation across items nested in dimensions and across followers (p) nested in leaders.
4, G (O1 Score)	$\frac{\sigma^2_{p:l}}{\sigma^2_{p:l} + \frac{\sigma^2_{pd:l}}{n_d} + \frac{\sigma^2_{pi:ld,e}}{n_{i:d}n_d}}$	Follower (p) focus with overall scores. Aggregation across items and dimensions means their effects contribute to error.
4, G (O2 Score)	$\frac{\sigma^2_{l}}{\sigma^2_{l} + \frac{\sigma^2_{p:l}}{n_{p:l}} + \frac{\sigma^2_{l:d}}{n_d} + \frac{\sigma^2_{li:d}}{n_{i:d}n_d} + \frac{\sigma^2_{pd:l}}{n_{p:l}n_d} + \frac{\sigma^2_{pi:ld,e}}{n_{p:l}n_{i:d}n_d}}$	Leader (l) focus with overall scores. Aggregation overall across followers (p), items, and dimensions.
4, Phi (O1 Score)	$\frac{\sigma^2_{p:l}}{\sigma^2_{p:l} + \frac{\sigma^2_{d}}{n_d} + \frac{\sigma^2_{i:d}}{n_{i:d}n_d} + \frac{\sigma^2_{pd:l}}{n_d} + \frac{\sigma^2_{pi:ld,e}}{n_{i:d}n_d}}$	Follower (p) focus. Aggregation overall across items and dimensions.
4, Phi (O2 Score)	$\frac{\sigma^2_{l}}{\sigma^2_{l} + \frac{\sigma^2_{p:l}}{n_{p:l}} + \frac{\sigma^2_{d}}{n_d} + \frac{\sigma^2_{i:d}}{n_{i:d}n_d} \frac{\sigma^2_{ld}}{n_d} + \frac{\sigma^2_{li:d}}{n_{i:d}n_d} + \frac{\sigma^2_{pd:l}}{n_{p:l}n_d} + \frac{\sigma^2_{pi:ld,e}}{n_{p:l}n_{i:d}n_d}}$	Leader focus. Aggregation overall across items, followers (p), items, and dimensions.

Note. Universe = universe score variance, error = measurement error variance. Generalizability (G) coefficients are specified for relative decisions and Phi coefficients are specified for absolute decisions. Total variance relevant to the equation always appears in the denominator. Therefore, each G and Phi coefficient represents the proportion of universe score variance to total variance. Focus = focus (or object) of measurement. p = participants or followers, i = rating items, l = leaders, d = dimensions, *e* = residual error, : = one level of nesting.