The Implications of Unconfounding Multisource Performance Ratings

Duncan J. R. Jackson
King's College London and University of the Western Cape

George Michaelides
University of East Anglia

Chris Dewberry
Birkbeck University of London

Benjamin Schwencke
Test Partnership Ltd

Simon Toms
Psychological Consultancy Ltd

Abstract

The multifaceted structure of multisource job performance ratings has been a subject of research and debate for over 30 years. However, progress in the field has been hampered by the confounding of effects relevant to the measurement design of multisource ratings and, as a consequence, the impact of ratee-, rater-, source-, and dimension-related effects on the reliability of multisource ratings remains unclear. In separate samples obtained from 2 different applications and measurement designs (N_1 [ratees] = 392, N_1 [raters] = 1495; N_2 [ratees] = 342, N_2 [raters] = 2636), we, for the first time, unconfounded all systematic effects commonly cited as being relevant to multisource ratings using a Bayesian generalizability theory approach. Our results suggest that the main contributors to the reliability of multisource ratings are source-related and general performance effects that are independent of dimension-related effects. In light of our findings, we discuss the interpretation and application of multisource ratings in organizational contexts.

The Implications of Unconfounding Multisource Performance Ratings

Job performance ratings have long held a pivotal role in industrial-organizational (I-O) psychology and are depended on as an important criterion in validation studies (Knapp, 2006), as a basis for guiding employment decisions (Borman & Motowidlo, 1997), and as a guide for developmental feedback (Toegel & Conger, 2003). Of appeal in I-O psychology is the multisource (or 360-degree) format, in which ratees are assessed by a number of different raters, each of whom approaches the rating task from one of several different possible role perspectives (referred to in the literature as "sources", e.g., managers, supervisors, and peers; see Borman, 1974). An advantage of the multisource approach to performance ratings is the richness of the information that is derived from this process (Mount, Judge, Scullen, Sytsma, & Hezlett, 1998). Such data-richness is the result of a complex arrangement of multiple, systematic influences on ratings assigned to an assessee (e.g. rating items, dimensions, different raters etc.). The complexity of this measurement design has fueled debate concerning the influence, status, and role of the different systematic sources of variance that could affect multisource ratings (e.g., Kraiger & Teachout, 1990; Lance, Hoffman, Gentry, & Baranik, 2008; Mount et al., 1998).

We argue that the true structure and reliability of multisource performance ratings has not yet been adequately addressed in the literature. Although I-O psychology researchers have made considerable progress towards identifying and clarifying the practical and statistical challenges associated with analyzing multisource data (e.g., B. J. Hoffman, Lance, Bynum, & Gentry, 2010; O'Neill, McLarnon, & Carswell, 2015), limitations in research design and analysis have prevented the discipline from progressing towards a complete understanding of the structure underlying the reliability of multisource ratings. The primary research gap here is that in previous studies, only limited subsets of relevant, systematic effects have been included for analysis, rather than a complete set of effects. In the absence of an estimation of all effects relevant to multisource ratings, the discipline is left with an incomplete representation of the measurement structure of this popular assessment approach. This has hampered our understanding of the measurement basis for multisource ratings and thus the development of theory and practice related to this approach. To address this limitation, we present analyses based on Bayesian generalizability theory[1] (G theory, see Brennan, 2001; Gelman, Carlin, Stern, & Rubin, 2013; LoPilato, Carter, & Wang, 2015) from datasets relating to two operational multisource rating procedures. We contribute to the research literature by estimating a comprehensive set of effects commonly cited as being central to multisource measurement designs. Moreover, we demonstrate that the structure of multisource ratings can only be unconfounded by simultaneously estimating a complete set of measurement-design-relevant effects, acknowledging the influence of aggregation, and considering generalization across different sources of error.

Reliability and Multisource Ratings

Multisource ratings present a popular option, often for the purposes of employee development: highlighting their importance for the performance growth of organizations (Bernardin, Konopaske, & Hagan, 2012; Lance, Baxter, & Mahan, 2006; Zimmerman, Mount, & Goff, 2008). Given the central role that job performance measures play in both research and practice in applied psychology and management, much attention has been directed towards their measurement characteristics. The measurement of job performance generally has been the subject of critical scrutiny since the dawn of modern psychology, with foundation researchers such as Thorndike (1920, p. 28) commenting that multiple raters in different roles were “unable to treat an individual as a compound of separate qualities”. Criticisms of a similar nature pervade more recent research, with Murphy (2008, p. 157) stating that the “relationship between job performance and ratings of job performance is likely to be weak”. Even more recently, LeBreton, Scherer, and James (2014, p. 482) asserted that in the light of the current, and typically low, estimates of their reliability, supervisory ratings of performance “appear to be fundamentally flawed”.

On this point, it is reliability that forms the foundation for all measurement. In the classic tradition of Spearman (1907, pp. 161-162), reliability is concerned with addressing “the observational

[1] Some of the concepts that we discuss as relating to “reliability” are often taken as evidence for “validity” in different contexts. For consistency with other studies where similar methodology has been employed, we maintain the traditional classification and present our effects as they relate to reliability. However, it has long been argued that generalizability theory “blurs the distinction between reliability and validity” (Brennan, 2000, p. 9; Cronbach, Gleser, Nanda, & Rajaratnam, 1972, p. 380).

errors and the irrelevant factors" involved in the measurement of a construct or in the estimation of the true relationship between two or more constructs. If a measure is *un*reliable, then it is neither possible to determine what is indicated by that measure nor to interpret, in any meaningful sense, its relationship with other measures (Viswesvaran, Ones, Schmidt, Le, & Oh, 2014).

Appropriate reliability estimation requires a consideration of measurement design characteristics and the popularity of multisource performance ratings is likely influenced by their comprehensive nature, based on the multiple perspectives on performance that are connected to their measurement design (Borman, 1974; Church & Bracken, 1997; Mount et al., 1998). The multifaceted structure of multisource performance ratings raises special considerations when evaluating their measurement characteristics. In particular, isolated estimates of agreement, interrater reliability, or internal consistency are insufficient in the context of multifaceted measurement designs (Cronbach et al., 1972; Cronbach, Rajaratnam, & Gleser, 1963). This is because there are multiple, interacting sources of variance relevant to multisource ratings that need to be considered simultaneously if the goal is to draw conclusions about their structure.

Multifaceted Perspectives on Reliability Estimates

Several researchers have presented a multifaceted perspective on multisource ratings, with published studies dating back around 3 decades. However, each of the relevant studies that we were able to identify showed evidence of methodological limitations concerned with the confounding of effects relevant to the measurement design of multisource ratings. Such confounding has arisen from data unavailability or from restrictions associated with specific statistical methods. In the following section, we summarize these limitations as they relate to studies invoking a multifaceted perspective on multisource ratings.

Findings suggesting large source-related effects. Borman (1974) suggests that source-related effects are an advantageous feature of multisource rating procedures because they allow for the representation of meaningfully distinct perspectives on performance. Conversely, and as discussed below, Mount et al. (1998) conceptualize such effects as being associated with unreliability. Regardless of perspective, source-related effects have represented a focus for researchers and have historically been assumed to arise from theoretical interactions between sources and dimension-based judgements (i.e., systematic differences between sources in the dimensions used to evaluate ratees, see Guion, 1965; Klimoski & London, 1974).

Table 1 presents results from a selection of studies of multisource performance measures in which multiple effect estimates were available. In one of the earliest of these studies, Kraiger and Teachout (1990) estimated source-related effects to be of a considerable magnitude. Similar findings have been replicated in other studies (B. J. Hoffman et al., 2010; Lance et al., 2008) and these are well-documented in reviews and texts on the topic (Aguinis, 2019; DeNisi & Murphy, 2017). However, whilst these studies indicate that source-related effects are of note, they report estimates that vary with regard to impact, with effect sizes ranging from small (7% of variance explained, see Scullen, Mount, & Goff, 2000) to large (around 34%, see Kraiger & Teachout). Given the complex nature of the measurement designs used in these studies, this variability is potentially explained by the confounding of multiple sources of variance: particularly in light of findings associated with confounding in related contexts (Jackson, Michaelides, Dewberry, & Kim, 2016; Putka & Hoffman, 2013).

Missing from the effects included in the Kraiger and Teachout (1990) design, but central to multisource ratings generally, are those related to raters. In the absence of information relating to raters, rater-related effects comprise a *hidden facet*[2] and are thus implicitly confounded with other effects. An important distinction in the context of multisource ratings relevant to the present discussion is that between rater- and source-related variance. The presence of discrete sources (e.g., managers, peers, etc) implies that multiple, individual raters are nested within each source. If, within sources, individual raters disagree or are unreliable in their assessment, estimates of "pure" group-level source effects are disrupted.

[2] We use the term "facet" to describe any systematic source of variance in a given measurement design (e.g., raters, sources, items) that is not the object of measurement (i.e., ratees in the present case). *Hidden* facets are those that are undeclared but are nonetheless relevant to a measurement model. Facets might be undeclared because of the availability of only 1 of potentially >1 levels of the facet, omission, or unavailability (see Brennan, 2001).

Put another way, if within-source (i.e., rater-related) variability is substantial, this could overwhelm any between-source variability. Thus, rater-based variance would, in these circumstances, contribute to error[3].

A similar potential limitation with respect to the confounding of source- and rater-related effects in the Kraiger and Teachout (1990) study was also apparent in a study by Guenole, Cockerill, Chamorro-Premuzic, and Smillie (2011). In the Guenole et al. study (see p. 207), restrictions around data availability meant that for the purposes of analysis, ratees were assessed by one representative sampled from each of 4 sources (including a self-, manager-, peer-, and direct-report-rating). While the authors reported large source-related effects, it is not possible to distinguish between source- and rater-related effects in this design. To achieve such a distinction, it would be necessary to include for analysis all of the raters in each of the sources under scrutiny (with the obvious exception of the self-rating). In the Guenole et al. study, because multiple raters were not nested within sources for analysis purposes, it is not possible to determine whether the effects in question are source-related, rater-related, or perhaps the result of some other systematic or unsystematic influence[4].

As a general principle related to multifaceted measurement designs, it is necessary to include multiple observations for any higher-order, group-level category to allow separation between different levels in an analysis. For example, if a multisource rating system includes several peers, a single supervisor, and a self-rating, it is possible to separate the effect for individual peers from the group-level peer effect. However, it will not be possible to achieve such a separation for supervisors in this case, because the individual supervisor effect here is the same as, and is thus confounded with, the group-level supervisor effect. Self-ratings, on the other hand, represent an exception where there is only ever the perspective of one individual about themselves. Thus, a separation between individual- and group-level effects is irrelevant to the self-rating.

[3] Error in this context refers to any contribution to unreliability in scores. Such sources of unreliability can be systematic (e.g., assessee × rater interactions) or residual in nature.

[4] There are other limitations associated with the sampling approach described here, which we cover below in the section on the B. J. Hoffman et al. (2010) study.

Findings suggesting large rater-related effects. In both the Greguras and Robie (1998) and Greguras, Robie, Schleicher, and Goff (2003) studies of multisource ratings, source-related effects were not estimated within the same analysis. While the authors collected information on different sources, the perspective of each source was presented in separate analyses. Thus, source-related effects were confounded in each of these analyses. This raises a contrasting limitation to that described above for the Kraiger and Teachout (1990) study with respect to confounding. Table 1 shows effects for one of the sources (i.e., supervisors) reported in Greguras and Robie[5], in which large effects were found for a rater-related effect and for a ratee main effect (i.e., a general performance effect). Similar results, although more detailed with respect to rater-related effects (see Table 1), were reported in O'Neill et al. (2015) where, again, source-related effects were not estimated. In the absence of source-related estimates in these studies, it is not possible to determine the influence that different sources have on rating variance. In a multisource rating design where source-related effects are omitted, uncertainty arises concerning whether rater-related variance should be interpreted as contributing to error or whether it should be interpreted as contributing to true score variance (e.g., see Murphy & DeShon, 2000) because the potentially true-score-relevant *source* variance is confounded with the unreliable *rater* variance.

In contrast to the studies mentioned above, Mount et al. (1998) acknowledged both rater- and source-based effects in multisource ratings. However, these effects were not estimated simultaneously in their study. Mount et al. found evidence that a model represented by 7 idiosyncratic rater factors and 3 dimension factors fit better than a comparison model represented by 4 source factors and 3 dimension factors. Based on the relative fit between these models, Mount et al. concluded that source-related effects are most appropriately considered as idiosyncratic rater-related sources of variance. However, the Mount et al. study conflated source-based effects in the former model and conflated rater-based effects in the latter model, creating challenges regarding interpretation (see B. J. Hoffman et al., 2010 for further discussion on this issue). To address this limitation and to extend the Mount et al. study, Scullen et al.

[5] For brevity, we only report results from Greguras and Robie (1998) in Table 1 because they were similar to the results reported in Greguras et al. (2003).

(2000) simultaneously addressed multiple effects, including those related to sources and those related to raters (see Table 1). Scullen et al. reported rater effects around 7 times larger than source-related effects. Yet, in this study, the authors applied a correlated uniqueness confirmatory factor analytic (CFA) parameterization as their basis for estimation. Researchers have since raised concerns about correlated uniqueness parameterizations in CFA, which have been shown to produce misleading results within the context of multifaceted measurement (see B. J. Hoffman et al., 2010; Lance, Lambert, Gewin, Lievens, & Conway, 2004; Lance, Noble, & Scullen, 2002). It is possible that such concerns might be relevant to the findings in the Scullen et al. paper.

To address the possible limitations of the Mount et al. (1998) and Scullen et al. (2000) studies, B. J. Hoffman et al. (2010) tested a CFA model that simultaneously included general performance, dimension-related, source-related, and rater-related effects. The authors tested separate factors for each of these effects. Hoffman et al. found a large rater-related effect (57.60%, see Table 1) and, contrary to the findings of Mount et al. and Scullen et al., they found that source-related effects also explained a substantial proportion of variance (22.00%).

While the B. J. Hoffman et al. (2010) study offered an advancement over preceding studies on multisource ratings, their methodological approach with respect to rater-related effects potentially introduced additional limitations. Specifically, raters in the Hoffman et al. study were not uniquely identified. To configure their data set such that it was amenable to traditional, CFA-based multitrait-multimethod analysis, Hoffman et al. randomly allocated 2 rater representatives for each rater source and coded them as *rater 1* and *rater 2* for each ratee. Although this approach creates a data structure that is amenable to analysis by CFA, it can lead to challenges regarding interpretability (see Putka, Lance, Le, & McCloy, 2011, p. 506). One reason for this is that the outcomes associated with a given random sampling of raters might not replicate across other such samples, resulting in uncertainty about which modeled solution is the correct one. Another reason is that an arbitrary approach to rater allocation akin to that described above does not account for the *ill-structured* nature of the design that is commonly employed in multisource ratings and in other multifaceted measures often applied in I-O psychology.

To clarify the issue of ill-structured designs in applied performance assessment, the assignment of >1 rater for each ratee is often found to be "ill-structured" in practice, such that raters are not generally the same for each ratee (i.e., the design is not fully crossed) but they are also not necessarily completely different for each ratee either (i.e., the design is not completely nested, see Putka, Le, McCloy, & Diaz, 2008). To illustrate with a simplified example, imagine a case where ratee A is assessed by rater 1 and rater 2. Imagine that ratee B, on the other hand, is assessed by rater 1 and rater 3. In this case, rater 1 is crossed with ratees. However, raters 2 and 3 are nested in ratees A and B respectively and thus the design is neither fully crossed nor completely nested and so is said to be ill-structured. Further challenges to the interpretation of the B. J. Hoffman et al. (2010) study arise because of a possible mismatch between the model tested and the true, ill-structured nature of the observed data set.

Multifaceted Effects and their Magnitudes Across Studies

Table 1 shows a cross-study comparison of multisource rating-related effects and their associated magnitudes published in widely-cited journals about which the above concerns related to confounding are relevant. Of the effects that are cross-study-analogous in Table 1, ratee or participant main effects (i.e., general performance effects) vary considerably from one study to the next. General performance effects indicate that some ratees are rated higher than others overall, regardless of any specific performance dimensions or source perspectives (e.g., Viswesvaran, Schmidt, & Ones, 2005). In B. J. Hoffman et al. (2010), the effect of general performance was very small (at 3.50%). Yet, in Greguras and Robie (1998), this effect was considerably larger (at 24.03%).

In addition to general performance, Table 1 shows cross-study-analogous effects relating to performance dimensions. Participant- (i.e., ratee) by-dimension-analogous effects imply that some ratees perform better on some dimensions than on others, regardless of general performance effects or source perspectives (e.g., Lance, Teachout, & Donnelly, 1992). Such dimension-related effects in Table 1 are somewhat more consistent across studies than are estimates for general performance and tend to explain relatively small proportions of variance (between 6.00% and 11.00%).

Another important consideration is that pertaining to source-related effects. Participant-by-source effects imply that a ratee's score depends on the perspective of group-level sources or role, regardless of general performance or dimensions (e.g., Borman, 1974). Such effects vary considerably in magnitude across the studies listed in Table 1 (between 7.00% and 33.74%), but this is possibly due to the confounding of rater-related effects as discussed above. On that note, participant-by-rater effects, indicating idiosyncratic rater influences (Mount et al., 1998), are consistently substantial across the studies listed in Table 1 (between 25.00% and 62.00% of variance). But again, it is difficult to draw firm conclusions here, given the potential for confounding with source-related effects in the literature.

Towards an Unconfounded Perspective on Multisource Performance Ratings

Our review of the literature suggests that to develop an unambiguous evaluation of the structure of multisource ratings, it is necessary to present a multifaceted perspective that controls for all effects relevant to the multisource measurement design. Implied here is that these effects should be acknowledged simultaneously so that each effect is controlled for the presence of all other effects. The studies shown in Table 1 and elsewhere in the literature are suggestive of a general measurement design for multisource ratings that typically consists of participant ratees (p) assessed by raters (r) nested within sources (s) on the basis of rating items (i) nested within performance dimensions (d).

In addition to a comprehensive acknowledgement of the multisource rating measurement design, there are two other reliability-related considerations that have yet to be fully addressed, but which could affect how variance is expressed in multisource rating procedures. The first of these considerations relates to aggregation. In practice, ratings from multisource procedures are likely to be aggregated across levels relating to specific facets. We suggest two different aggregation scenarios that are potentially relevant to multisource ratings, as follows: (a) rating items aggregated to form dimension scores and (b) dimension scores aggregated across different raters within each source. Other types of aggregation across sources are possible, but, as discussed later, could be difficult to defend given the potential magnitude and meaningfulness of source-related effects. It is possible that each of the two aggregation configurations described above could result in different outcomes regarding variance explained by the various effects in the measurement model. Evidence from the literature on assessment centers suggests that aggregation can affect the magnitude of variance components (Jackson et al., 2016; Kuncel & Sackett, 2012; Putka & Hoffman, 2013). Yet, to our knowledge, the influence of aggregation has not been studied in the context of multisource ratings.

The second reliability-related consideration that has yet to be fully investigated in the multisource performance literature concerns intentions to "generalize" across different sources of error (see Cronbach et al., 1963). From a G theory perspective, reliability is not simply a property that is inherent to a set of scores, but it is dependent on how the scores will be used and interpreted. Different levels of reliability will result from different applications and interpretations of a given set of scores. The G theory perspective on reliability proposes that one or more effects might contribute to universe score (analogous to true score) variance, error, or to neither of these categories (Brennan, 2001; Shavelson & Webb, 1991). If an effect is deemed to contribute to error, then it is considered as a facet across which the researcher wishes to generalize the universe-score-related elements of their measurement design.

For example, consider an organization that employs a multisource rater system and uses the source-level perspective on performance to provide developmental feedback. In this case, individual rater-related variance is deemed to contribute to error and rating sources are deemed to contribute to universe score variability. This is because if different raters vary in their judgment within sources, then it will interfere with the source-level perspective. The researcher's intention, therefore, is to generalize source-related effects across rater-related effects to arrive at a meaningful score. G theory allows the researcher the flexibility to specify variance sources as contributing to error or to universe score variance, depending on how they will be used. With respect to multisource rating measurement designs, it is possible that researchers will be interested in the reliability-related outcomes associated with generalizing across different items, raters, or, perhaps in some circumstances, rating sources (e.g., if across-source agreement is desired), and relevant combinations of these design features.

Putka and Hoffman (2014) addressed aggregation as it pertains to performance assessment in their methods chapter, which provides a methodological overview of G theory applied to multisource rating data. Putka and Hoffman re-analyzed data from Greguras and Robie (1998) to demonstrate that different types of generalization and measurement design can result in different reliability-related outcomes. However, their analysis was restricted by the omission of source effects and other confounded estimates inherent in the original Greguras and Robie (1998) study. Putka and Hoffman addressed this issue by proportionately dividing confounded variance components based on previous findings from the literature. This approach was entirely adequate for the purposes of a methodological overview. However, it was not intended to inform on the unconfounded reliability of multisource measures. It also could not inform the literature in this respect because of the confounding observed in the original Greguras and Robie study and, as our review highlights, in previous studies that informed Putka and Hoffman's approach to dividing variance estimates. Thus, to our knowledge, different intentions regarding generalizability have not yet been presented from an unconfounded perspective in the multisource ratings literature.

Summary

A multifaceted perspective on multisource ratings has been tested in different forms in the research literature. However, the interpretability of findings related to this line of research is limited by the confounding of one or more features that are central to the measurement design of multisource ratings. To address this limitation, we adopt an approach to the analysis of multisource ratings that (a) simultaneously acknowledges the systematic sources of variance chiefly relevant to their measurement design, (b) acknowledges the effects of aggregation, and (c) acknowledges different intentions with respect to generalization. By providing unconfounded estimates of the impact of a comprehensive set of effects, we seek to add clarity to the literature. In addition, this new perspective can provide applied researchers and practitioners with valuable information about the likely consequences of generalizing multisource ratings in different ways (e.g., across raters) on the reliability of their aggregated rating data.

We propose three Research Questions related to these aims, as follows:

Research Question 1: What proportion of the variance in multisource ratings is uniquely associated with each of the following: ratees, raters, sources, items, dimensions, and their interactions?

Research Question 2: To what extent does the reliability of multisource ratings depend on the approach used to aggregate them (e.g., aggregation across items to form dimension scores)?

Research Question 3: To what extent does the reliability of multisource ratings depend on generalization intentions (e.g., generalization across different raters or generalization across rating items and raters)?

Method

We analyzed data from two operational multisource rating procedures, each of which was from different organizations and represented different measurement designs. Details relating to these samples are provided below.

Sample 1

Participants. Participants in Sample 1 included 392 managerial ratees (298 males, 94 females) and 1495 unique raters (1121 males, 374 females) from a manufacturing organization based in the United Kingdom. Groups of raters were nested in sources (except for self-ratings, which were treated as a source-level effect only), including representatives from senior manager, colleague, direct report, self, and stakeholder roles. The multisource rating procedure in Sample 1 was used primarily for employee development. Data on ethnicity and age were not collected by the organization in Sample 1 due to confidentiality concerns.

Measurement design. The measurement design in Sample 1 was configured such that all participant ratees (p) were assessed by raters (r, on average 2 per source) who were nested in sources (s, on average 5 per ratee) on the basis of rating items (i, on average[6] 16.46 per dimension), which were nested in performance dimensions (d, a total of 4). This represents the “classic” design often described in the multisource rating literature (for other examples of this design, see Table 1).

[6] In keeping with Brennan (2001), we applied harmonic mean values to items.

Sample 2

Participants. Participants in Sample 2 included 342 managerial assessees (216 males, 126 females) with a mean age of 38.31 (SD = 9.65) and 2636 unique raters (1579 males, 1057 females) with a mean age of 40.24 (SD = 9.89). The multisource rating system in Sample 2 was available for use by different organizations based in the United Kingdom in the banking, retail, accounting, insurance, human resources, and management consulting industries. The purpose of the assessment in Sample 2 was for a mixture of performance assessment and development functions, depending on client requirements. Groups of raters were nested in sources, including representatives from senior manager, peer, direct report, and client roles, with self-ratings treated as a source-level effect.

Measurement design. Sample 2 was configured such that all participant ratees (p) were assessed by raters (r, on average 2 per source) who were nested in sources (s, on average 5 per ratee) on the basis of rating items (i, on average 10.04 per dimension), which were nested in performance dimensions (d, a total of 24). Between 4 to 6 (*M* = 5.40, *SD* = .89) dimensions were, in turn, nested within each of 5 broad dimension categories (c). Broad dimensions have been explored elsewhere in the I-O psychology literature and have been applied in the assessment center context (see B. J. Hoffman, Melchers, Blair, Kleinmann, & Ladd, 2011). The nature of the measurement design in Sample 2 was such that different clients accessed the assessment system and therefore some variation was apparent in the number of levels associated with specific sources of variance.

Development of Multisource Rating Procedures in Samples 1 and 2

Both multisource rating procedures, their associated rating items, dimensions, and rater training procedures were developed and conducted in accordance with relevant professional and academic guidelines (e.g., Committee on Test Standards, 2011; Pulakos, 1986) and were based on job analyses of the focal positions (e.g., Williams & Crafts, 1997). Example items for Samples 1 and 2 respectively included: *Ensures the strategy, objectives and activities of the team are focused on addressing customer needs* and *Gives ongoing and constructive performance-related feedback*. The scale in Sample 1 ranged from 1 (the rater has never observed this behavior) to 5 (the rater always observes this behavior). The scale in Sample 2 was a normed scale ranging from 0 to 100 that was based on an original scale ranging from 1 (strongly disagree) to 6 (strongly agree).

Training was provided to those who administered and provided feedback on each rating process. In Sample 1, this involved attending a formalized workshop, an introduction to the process used, use of the associated online platform, and experiential learning based on mock assessments coupled with a comparative evaluation of ratings assigned by assessors (akin to frame-of-reference training, see Pulakos, 1986). In Sample 2, training involved completion of a face-to-face, half-day course covering procedural content, administration of the rating procedure, and the practice of rating in a similar manner to that described for Sample 1. The extent to which training was successful was not assessed in Sample 1. In Sample 2, training performance was assessed by a trainer at the conclusion of the program. While these training performance ratings were unavailable, it was mandatory for trainees to acceptably complete training. Performance evaluations were repeated annually in the Sample 1 organization, although we only had access to one of those evaluations in the present study. In Sample 2, the evaluation was primarily treated as a one-off assessment.

The participant organization in Sample 2 had conducted background exploratory factor analytic work to aid in the development of their broad dimension framework. Dimension definitions are provided in Table A1 of the Appendix. Rating scales across both samples described above were similar in structure to those described in the extant literature on performance ratings (e.g., Bennett, Lance, & Woehr, 2006) and raters were required to assess the proportion of time that they had observed the behavior described in each rating item. Each rating item related back to job-critical knowledge, skills, abilities, and other characteristics identified in the job analysis for each sample.

Data Analysis

Effects, measurement design, and generalization. The measurement designs described above for each sample resulted in a total of 17 separate effects that required estimation for Sample 1 and a total of 19 separate effects for Sample 2. For each sample, we were able to simultaneously estimate separate effects attributable to general performance (i.e., participant or ratee main effects), participant × dimension (i.e., dimension-related) effects, participant × source (i.e., source-related) effects, and multiple item- and rater-related effects. The presence of sources implies a measurement design whereby raters are nested in sources, which was the case in both of our samples. Because of the structure of this design, rater-related effects were always treated as contributing to error. In keeping with our research aims about generalization, we tested reliability-related outcomes resulting from treating as error: (a) sources and raters, (b) raters, (c) items and raters, and (d) a combination of items, sources, and raters. These four generalization types constituted all feasible generalization scenarios available from our data sets.

Bayesian inference. We used Bayesian inference as a basis for our G theory models, which has been applied previously in the wider literature on multifaceted assessment (e.g., Jackson et al., 2016; LoPilato et al., 2015). There have been recent calls to extend applications of Bayesian methods into other areas of scrutiny in the organizational context (Kruschke, Aguinis, & Joo, 2012; Zyphur, 2009; Zyphur, Oswald, & Rupp, 2015), to which we respond directly in this study. In the application of G theory methods, Bayesian inference is amenable to handling complex designs (i.e., designs with a large number of parameters) that might be computationally impractical with traditional, restricted maximum likelihood- (REML-) based estimation (Brennan, 2001). Moreover, REML-based estimators can return problematic, zero-fenced outcomes (Searle, Casella, & McCulloch, 2006), which can raise concerns about interpretability where fenced outcomes are evident.

Ill-structured measurement designs and aggregation. Raters and ratees were neither fully crossed nor nested in either of the samples in our study and thus both measurement designs were ill-structured in nature (see Putka et al., 2011). We therefore configured a hierarchical Bayesian model in order to contend with data sparseness (see Gelman & Hill, 2007). This permitted the analysis of ill-structured data sets without the need to delete large portions of data to arrive at a crossed design. We rescaled any rater-related variance and reliability estimates using the q-multiplier approach detailed in Putka et al. (2008). Formulae relating to our application of the q-multiplier are shown in the section that follows. Guided by formulae provided in the literature on G theory (e.g., Brennan, 2001; Putka & Hoffman, 2014), we rescaled our variance estimates so as to acknowledge the effects of aggregation to (a) dimensions and (b) dimensions aggregated across raters within sources[7]. We integrated these aggregation formulae into our Bayesian model to obtain samples from posterior distributions for all estimates.

Model specification. Our analyses were conducted using R 3.4.1 (R Core Team, 2017), Stan 2.17.0 (Stan Development Team, 2017), and Rstan 2.17.3 (Stan Development Team, 2018). Stan uses a Hamiltonian Monte Carlo sampling approach as a basis for Monte Carlo Markov Chain estimation (M. D. Hoffman & Gelman, 2014). For Sample 1, a hierarchical model was configured with 15 error terms and 1 fixed intercept. For Sample 2, an equivalent model was configured, but for 19 error terms and 1 fixed intercept, given the additional effects that required estimation in that sample. To facilitate cross-sample comparability, we scaled each raw dataset to 1 standard deviation. Recent findings challenge the tenability of normality assumptions applied to the measurement of performance (Aguinis, Ji, & Joo, 2018). Accordingly, we evaluated normality by inspecting density and QQ plots: neither of which suggested concerns arising from appreciable deviations from normality in either of our datasets. To both samples, we applied a normal prior to the fixed intercept with a mean of 0 and a standard deviation of 5. This constitutes a broad, weakly-informative prior that would permit convergence towards the grand mean of the data set.

The weakly-informative prior used here contains information sufficient to rule out values that are unreasonable (e.g., negative variance estimates) but is not restrictive to the extent that it excludes values that could possibly occur: even if that possibility is small. Thus, a prior of this nature can accommodate extreme values if they were to arise from the observed data-set. Moreover, our application of a weakly-informative prior has been utilized previously in a study of assessment center ratings (Jackson et al., 2016). We applied a non-centered reparametization to the model (Papaspiliopoulos, Roberts, & Sköld, 2007), which requires that random intercepts are sampled from the unit normal distribution which are, in turn, rescaled by multiplying them by group-level standard deviations for each random intercept. The prior distribution for these group-level standard deviations was the half-normal distribution with a

[7] We also considered aggregation involving (a) dimensions across raters in sources and across sources and (b) overall ratings across all items, dimensions, raters, and sources. However, these aggregation types were not necessarily defensible, given the large and potentially meaningful source effects that we observed (discussed later).

location of 0 and a scale of 1 standard deviation. Although the half-Cauchy distribution is often recommended for variance components models (Gelman, 2006; Jackson et al., 2016), the more informative, half-normal distribution is preferable here because of the presence of scaled data.

For each model we conducted the simulation with four chains and 10,000 iterations within each chain. Discarding the first 5000 iterations as warm-up, we used the remaining iterations for the analysis. Both models showed good convergence. Specifically, visual inspection of trace, density, and autocorrelation plots suggested good mixing of the chains without any autocorrelation issues whilst the scale reduction factor, effective sample size, and estimates of the Monte Carlo standard errors were acceptable for all model parameters (see Gelman & Rubin, 1992).

Results

Sample 1

Variance estimates for Sample 1. Table 2 shows variance estimates from Sample 1 for pre-aggregated ratings, items aggregated to dimensions, and dimensions aggregated across raters in sources (see Research Questions 1 and 2). Notable sources of variance at the disaggregated level included the main effect for participant ratees (σ_p^2, 20.39% of between-participant variance explained) and source-related variance (σ_{ps}^2 and σ_s^2, 19.17% and 20.60% respectively). Item-related ($\sigma_{pi:d}^2$) and residual variance ($\sigma_{pir:ds,e}^2$) were also of somewhat sizable magnitudes, which diminished once aggregation was accounted for. When aggregating to the dimension level (i.e., when aggregating across items to arrive at dimension scores), large portions of variance were attributable to participant ratee main effects (σ_p^2, 27.23%) and source-related effects (σ_{ps}^2, σ_s^2, and $\sigma_{pr:s}^2$: 25.60%, 27.51%, and 8.18% respectively). A similar pattern emerged when aggregating across raters in sources, with participant ratee main effects and source-related effects explaining large portions of variance whilst participant-by-dimension effects in Table 2 were very small, regardless of aggregation type (< 1.27%).

Reliability estimates for Sample 1. Table 3 presents reliability estimates corresponding to the variance estimates reported in Table 2 for Sample 1. The composition of effects that are regarded as contributing to universe score versus those contributing to error depends on the intentions of the researcher and thus the desired type of generalization (see Research Question 3). For example, if the researcher wishes to generalize across different raters, then any variance component associated with raters is treated as a source of error. Rater-related effects are often regarded as contributing to error because unreliability or disagreement among raters potentially interferes with source-level perspectives and/or summary scores. However, with source-related effects, it is possible to argue that sources contribute to universe score variance or error, depending on the intention of the practitioner or researcher. If the intention is to use the information from different sources in decision-making processes, then sources would be specified as contributing to universe score variance. Conversely, a desire to minimize variation among sources indicates that source-related variance should be treated as contributing to error. This might be due to a focus, for example, on dimension scores rather than source perspectives as a guide to decision-making.

The composition of universe score versus error variance for generalization across (a) sources and raters, (b) raters, (c) items and raters, and (d) items, sources, and raters was repeated for each type of aggregation level in Table 3. Sources of variance were adjusted for each aggregation level, also as shown in Tables 2 and 3. For example, when the desire was to aggregate to dimensions, the relevant formula acknowledged aggregations of items within each dimension.

The pattern of results in Table 3 shows that a decrement in reliability occurred when sources were considered to contribute to error. Therefore, a consideration of source-related variance as error was detrimental to reliability regardless of aggregation level (i.e., at the dimensions- and dimensions-across-raters-in-sources levels of aggregation, where reliability was ≤ .34). Conversely, when source-related variance was considered to contribute to universe score variance, the reliability-related outcomes were favorable (where aggregated ratings resulted in reliabilities ≥ .81).

Sample 2

Variance estimates for Sample 2. Table 4 shows variance estimates for Sample 2 relating to pre-aggregated ratings, ratings aggregated to dimension scores, and dimensions aggregated across raters in sources (see Research Questions 1 and 2). The measurement design for Sample 2 included a feature that was not present in Sample 1, in that dimensions were nested in broad dimension categories. Perusal of the disaggregated ratings relevant to between-participant variance in Table 4 reveals relatively sizable proportions of variance associated with participant ratee main effects (σ_p^2 = 19.59%), source-related variance (σ_{ps}^2 = 9.85%), and an interaction involving items, dimensions, and categories ($\sigma_{pi:d:c}^2$ = 33.85%). The participant-by-dimension-related effect at the disaggregated level was small ($\sigma_{pd:c}^2$ = 1.45%).

Consistent with the findings observed in Sample 1, Table 4 shows that aggregation to dimension scores revealed large participant ratee main effects (σ_p^2 = 38.23%), source-related effects (σ_s^2 = 11.68%; σ_{ps}^2 = 19.22%), with the addition of a moderate effect associated with broad dimension categories (σ_{pc}^2 = 9.81%). The profile when aggregating dimension scores across raters in sources was almost identical to that for aggregation to dimension scores only and reflected mainly participant ratee main effects, source-related effects, and a moderate participant-by-broad-dimension category interaction.

Reliability estimates for Sample 2. Table 5 shows reliability estimates corresponding to the variances reported in Table 4 for Sample 2. The pattern of results in Sample 2, vis-à-vis reliability, was generally consistent with the results for Sample 1. Across both types of aggregation (see Research Question 3), consideration of source-related effects as error led to a decrement in reliability. When aggregating to dimensions and to dimensions and raters in sources (see Table 5), treating sources as a component of error variance led to estimated reliabilities ≤ .60. Conversely, a consideration of sources as contributing to universe score variance led to increases in reliability, with estimates ≥ .84.

To provide an additional perspective on the reliability-related impact of the effects chiefly concerned with dimensions and broad dimensions in Sample 2 (i.e., $\sigma_{pd:c}^2$ and σ_{pc}^2, respectively), we estimated reliabilities across all levels of aggregation with both $\sigma_{pd:c}^2$ and σ_{pc}^2 removed from the reliability estimates that appear in Table 4 to determine the extent to which dimensions and broad dimensions, collectively, influenced reliability outcomes. We then compared the resulting reliability estimates with the original estimates that included the effects $\sigma_{pd:c}^2$ and σ_{pc}^2. The results of this comparison revealed minimal absolute differences (ranging from .01 to .07) between reliabilities when both $\sigma_{pd:c}^{2}$ and σ_{pc}^{2} were included versus when they were excluded from reliability equations across all levels of aggregation considered in this study. Thus, the cumulative impact of both specific and broad dimensions on reliability was minimal, irrespective of aggregation type.

Figures 1 and 2 show reliability point estimates and associated 95% credible intervals for all types of aggregation and generalization across both samples. Credible intervals represent the most probable values that a reliability parameter can take, in the present case, within 95% certainty (Gelman et al., 2013). At aggregated levels, Figures 1 and 2 suggest a much greater level of uncertainty when sources are specified as contributing to error. Moreover, Figures 1 and 2 clearly show a reduction in reliability in most cases when sources are considered to contribute to error.

Discussion

The literature on multisource job performance ratings has yielded inconsistent estimates of the extent to which rating scores reflect variance in general performance, dimension-, source-, and rater-related influences (see Table 1). These inconsistencies may be explained by an omission of one or more sources of variance relevant to multisource ratings. In the absence of relevant effect estimates, results are subject to the distorting influence of confounds, such as the confounding of source- and rater-related effects (e.g., Greguras & Robie, 1998; Greguras et al., 2003; Guenole et al., 2011; Kraiger & Teachout, 1990; O'Neill et al., 2015). In other studies, methodological concerns limit the interpretability of reported results (e.g., B. J. Hoffman et al., 2010; Mount et al., 1998; Scullen et al., 2000).

We sought to contribute to the multisource rating literature by addressing these issues directly. Specifically, we offer the first comprehensive, multifaceted, and unconfounded analysis of multisource ratings relevant to measurement designs often cited in the literature. The output of this analysis, reported here, is the first set of variance estimates free from the influence of between-effect confounds. Consistent across the 2 samples in our study, we found that the unconfounded structure of multisource ratings reflects source effects, general performance, and no other appreciable contributors to reliability. Below, we discuss, in turn, findings associated with sources, general performance, dimensions, and raters. We then examine the implications of these findings for researchers and practitioners.

The Profile of Unconfounded Effects in Multisource Ratings

Source effects. Source effects are well-documented in the literature on performance measurement (Aguinis, 2019; DeNisi & Murphy, 2017). Yet, a review of individual studies reveals considerable variability in the estimated magnitude of such effects. Table 1 shows a selection of studies of performance ratings and source-related effects that range from between 7% (Scullen et al., 2000) to around 34% of variance explained (Kraiger & Teachout, 1990). In response to Research Questions 1 and 2, our unconfounded results suggest that at both the dimension- and dimension-across-raters-in-sources levels of aggregation, large portions of variance in multisource ratings are attributable to source-related effects (i.e., σ_s^2 and σ_{ps}^2), which cumulatively explained between 30.90% and 58.06% of the variance in ratings across our samples. The finding of large source-related effects in our study is contrary to those in some previous studies in this context, which were possibly limited by confounding or methodological concerns discussed earlier (e.g., Mount et al., 1998; Scullen et al., 2000). Large source-related effects suggest that an important characteristic of multisource ratings is based on the perspectives of different roles on the performance of ratees. The role or source perspective in our study was separated from the rater perspective: the latter of which is best considered as a source of error in many or most cases.

In addition to providing the first estimate of the influence of source effects on performance ratings free from the influence of confounding, our analysis also casts light on the reason that these effects occur. The theoretical perspective offered by Guion (1965, pp. 472-473) suggests that source variation may occur because sources systematically differ in the dimensions they use to evaluate people: a proposition which found some support in a later study of performance ratings (Klimoski & London, 1974). However, in the present study, we found dimension-related effects to be trivial with relation to their magnitude (see the section on dimension effects below). Both participant × dimension interactions and participant × dimension × source interactions (or their analogues in Sample 2) were small. The former small effect here implies that the ratee evaluation does not depend on dimensions. The latter small effect implies that the source perspective does not depend on dimension-based evaluations. In the absence of meaningful rating variance associated with dimensions, Guion's suggestion that source effects are due to the use of systematically different dimensions across different sources is unpersuasive. More likely is the possibility that source effects arise because different sources are exposed to different types of information about an employee's behavior and that they interpret that behavior as it relates to their perspective (e.g., from the peer's perspective, the ratee tends not to engage socially; but from the manager's perspective, they tend to work diligently, see Borman, 1974; Borman & Motowidlo, 1997; Norman & Goldberg, 1966). Evidence for this is suggested by our findings of strong source and participant × source effects (see Tables 2 and 4).

General performance effects. Our findings related to large source effects were consistent with those of B. J. Hoffman et al. (2010). However, contrary to their finding of small (3.50% of variance explained) general performance effects, we found evidence of large general performance effects (σ_p^2, which explained between 27.23% and 39.22% of the variance in ratings across our samples). To a less extreme extent, our results were also of a higher magnitude with respect to general performance effects than the other studies shown in Table 1. Our findings suggest that an assessment of general performance represents a key property of the measurement characteristics of multisource ratings.

A general performance factor implies that some assessees consistently outperform others, regardless of any rating items, dimensions, or source perspectives. This could be a result of the *positive manifold* often observed across different contexts relevant to I-O psychology (Ree, Carretta, & Teachout, 2015). General consistency in behavior is suggestive of the influence of relatively stable, underlying traits (e.g., general mental ability, conscientiousness), which have been found in previous research to correlate with performance ratings (Salgado et al., 2003; Salgado, Anderson, & Tauriz, 2015). Such trait-based or behavioral factors could possibly be subsumed into a general performance factor (e.g., Putka & Hoffman, 2013). This represents an avenue for future research focusing on the extent to which general mental ability and conscientiousness covariates influence the magnitude of general performance effects in performance ratings. The trait-based or behavioral influences referred to here are distinct from those intended for assessment in multisource ratings in the form of behavioral performance dimensions.

Dimension effects. Notwithstanding concerns about confounding and methodology, one of the more consistent findings in the multisource rating literature is that dimension-related effects tend to be small-to-moderate in size (see Table 1). Our results were no exception in this respect, and the effects most centrally associated with dimensions (σ^2_{pd} and $\sigma^2_{pd:c}$) were very small across our two samples (between 1.15% and 2.91% of variance explained). This finding is consistent with results related to dimensions in other contexts in I-O psychology (e.g., Jackson et al., 2016; Lance et al., 2004; Putka & Hoffman, 2013). However, in our Sample 2, part of the measurement design included a set of broad dimensions into which regular dimensions were nested: akin to second-order dimensions discussed elsewhere (B. J. Hoffman et al., 2011). These broad dimension categories (σ^2_{pc}), whilst still moderate in magnitude (between 9.81% and 10.07%), explained more variance than that explained by regular dimensions. To explore this effect further, we tested the extent to which regular dimension- and broad dimension-related effects *collectively* (i.e., $\sigma^2_{pd:c}$+ σ^2_{pc}) contributed to reliability in Sample 2. This contribution turned out to be small. Specifically, regular and broad dimension-related effects ranged from between .01 to .07 in terms of their proportional contribution to universe score variance. Thus, our results ultimately suggest that the reliability of multisource ratings is not appreciably influenced by the presence of either regular dimensions or broad dimensions and their related effects.

Rater effects. Irrespective of their magnitude, source-related, general performance, and dimension-related effects could, at least in theory, contribute to universe score variance in multisource ratings, depending on the desired type of generalization. Although some debate is evident in the multisource rating literature (e.g., Murphy & DeShon, 2000), we argue, as others have (e.g., B. J. Hoffman et al., 2010), that rater-related effects ordinarily contribute to error. In our Sample 1 (see Table 2), aggregation to dimensions resulted in rater-related effects that collectively explained a substantial 16.91% of the variance in ratings. Aggregating dimensions across raters in sources reduced this influence in Sample 1, where the same rater-related effects collectively explained 9.16% of the variance. In Sample 2 (see Table 4), rater-related effects were generally small when aggregating to dimensions (6.35%) and even smaller when aggregating dimensions across raters in sources (4.23%). Thus, and in contrast to other perspectives proposed in the literature (Mount et al., 1998; Scullen et al., 2000), when rater-related effects are unconfounded from source-related effects, our results suggest that their influence on reliability is likely to be fairly minimal when aggregating across raters.

Implications for Researchers and Practitioners

Output-wise, what does a multisource rating instrument provide for researchers and practitioners? Our results suggest that the output from multisource ratings consists primarily of source-related perspectives, a general performance perspective, and little else. The source-related perspectives that we identified indicate that an employee's rated performance is influenced to a considerable extent by whether the person rating him or her is a manager, peer, or subordinate. For example, a senior manager might form an impression about an employee that differs from their peer's impression. It is possible that this impression could be based on that senior manager's background perspective (e.g., they consistently observed the employee working diligently and therefore assigned a high rating). The manager's perspective here might differ meaningfully from the peer's perspective (e.g., the employee did not engage in social activities and was therefore assigned a lower rating by their peer group).

The general performance perspective, on the other hand, could be related to underlying traits (e.g., general mental ability or personality) of participant ratees because such characteristics can, in theory, be subsumed into general performance factors (e.g., Putka & Hoffman, 2013; Ree et al., 2015). It might be the case that raters are able to detect such individual differences and that ratees are provided with opportunities to express underlying dispositional characteristics. Our results suggest that these dispositional characteristics are not the dimensions intended for assessment. We found that when dimension-related effects were removed, there was no appreciable effect on reliability estimates. Put another way, dimension-related effects were so small that for practical purposes they were irrelevant to reliability in the multisource rating procedures under scrutiny.

Theoretical perspectives have historically posited that dimension-based judgments are dependent on sources, in that hypothetical systematic differences exist in way that different sources use dimensions (Guion, 1965; Klimoski & London, 1974). For example, it might be that managers place a greater weighting on performance-related dimensions, whereas peers might place a greater weighting on interpersonally-oriented dimensions. This perspective implies that a 3-way interaction should be evident involving participant ratees × dimensions × sources. However, when we tested this effect, it explained a negligible (< 2%) proportion of variance across our samples. Our evidence suggests that sources (i.e., managers, peers) essentially bypass specific dimensions altogether and form source-dependent overall impressions of ratees. Evidence for this is observed in the 2-way interaction involving participant ratees × sources and source main effects in an ill-structured design, which separately and collectively explained substantial proportions of variance, as discussed previously. Dimensions have no bearing on this type of evaluation and thus our findings raise questions about the role of dimensions in theoretical frameworks relating to multisource ratings. A suggested revision to extant theoretical perspectives is that source perspectives could be based on an overall impression of ratees and not on source-dependent dimension judgments.

Given the evidence that has accumulated on the small effects associated with dimensions here (see Table 1) and in other contexts (e.g., Lance, 2008), it is possible, as was identified almost a century ago, that raters continue to be “unable to treat an individual as a compound of separate qualities” (Thorndike, 1920, p. 28): at least in terms of the qualities purported in dimension frameworks. Nevertheless, our results suggest that raters might be able to offer valuable source-specific and general perspectives on performance. Such a conclusion has weighty implications for the architects of multisource job performance ratings.

In applied settings, it might be the case that multisource rating procedures are developed with the intention to generate dimension scores that can be used for developmental guidance or for promotion decisions (e.g., an average score for communication skills, teamwork, etc.). However, our results suggest that summary dimension scores do not represent the structure of multisource ratings and are therefore not a meaningful approach towards summarizing multisource performance ratings. If practitioners were, in contrast, to use the structure suggested by our findings, then they might create summary scores for each source and for overall, general performance. For example, an average score could be generated for the manager impression, a separate score for the peer impression, a separate score for the client impression, etc. A grand mean score across all ratings could be used as an indication of general performance. An important distinction here is that specific dimensions, as conceptualized in competency frameworks, are not involved in the generation of source-based scores because our findings suggest that they are ultimately not involved in the structure of multisource performance ratings.

An appropriate approach to aggregation is another, related point for practitioners to consider and our results suggest that the choice of aggregation type could affect the reliability of the assessment. Because almost all job performance measurement designs are multifaceted, it is unlikely that there will ever be a single, relevant estimate of reliability for this type of measure. Due to their multifaceted nature, reliability in performance ratings depends on how such ratings are aggregated and the measurement conditions (e.g., raters, items) over which the researcher or practitioner wishes to generalize (Cronbach et al., 1972). Given the magnitude of source-related effects in our study, that they are likely to be meaningful (Borman, 1974), and the detrimental effects on reliability when they are considered to contribute to error (see Figures 1 and 2), we suggest that it is only defensible to consider aggregation across rating items and across raters nested in sources: but not across sources. If the intention was to generalize across sources, then our estimates suggest that the reliability of multisource ratings could be as low as .31 (in Sample 1) or .52 (in Sample 2). A more defensible approach would be to conceptualize source-related effects as though they contribute to universe score variance (see Research Question 3). Under this perspective, a conservative evaluation of our results suggests that the reliability of multisource ratings could be much more encouraging and in the range of between .81 (in Sample 1, see Table 3) and .84 (in Sample 2, see Table 5).

Limitations and Areas for Future Research

One aim of this study was to demonstrate that by unconfounding the systematic sources of variance in multisource ratings in two different studies, the results derived from these studies would be consistent. This aim was largely achieved. However, there are several limitations that deserve consideration. Firstly, our approach to data analysis using G theory with Bayesian inference was not the only approach that we could have taken. CFA models are also capable of handling data sets of this nature and can provide more detail about specific model characteristics than can G theory models (Le, Schmidt., Harter, & Lauver, 2010; Putka et al., 2011). Nonetheless, given the size of our models and therefore the number of parameters that would require estimation under a CFA framework, a Bayesian G theory approach offered a practical alternative and one that has been successfully applied in other contexts in I-O psychology (LoPilato et al., 2015; Zyphur et al., 2015). It also provided a level of detail sufficient for us to address our research aims.

Secondly, we included two, sizable samples, which were intended to be similar enough to allow comparability, but sufficiently different to present a reasonable evaluation of cross-sample effects. One of our samples was from a specific organization and the other included representatives from multiple organizations. It would be beneficial to the discipline to extend our findings beyond the samples included in our study and into different organizational contexts and different organizational levels. On this note, it would also be of interest to examine whether results similar to ours generalize to other measurement designs. One design feature that we did not explore was that related to occasions (see Brennan, 2001; Le, Schmidt, & Putka, 2009 for a discussion about the occasions facet), because this is not typically described as a key feature of the multisource measurement design in the literature (see Table 1). For organizations with an ongoing employee development program and a relatively stable employee population, occasions might represent an effect of interest. In this respect, interactions involving participant ratees × raters × occasions might be of consequence as a source of error. It is possible to accommodate a within-participants design feature such as this into a Bayesian G theory model and we hope that this idea will be explored in future studies.

Thirdly, we did not investigate the role of personality and mental ability and the nature of its relationship with performance ratings because of our focus on the measurement design internal to multisource measures. As discussed above, we suggest that personality, particularly other-rated personality (see Connelly & Ones, 2010), and mental ability might have a bearing on the general performance factor in multisource ratings. An investigation into this possibility could be achieved using several different approaches, including augmenting a variance components model or a CFA model with a set of trait-based covariates. This represents an important avenue for future research into the psychological factors underpinning the operation of multisource ratings. Findings related to a study of this nature could help to inform theory related to multisource ratings that could, in turn, influence new research directions.

Conclusion

The reliability of multisource performance ratings is a topic that has been debated in the literature and with good reason, given the crucial role that performance measures play in applied psychology (LeBreton et al., 2014; Murphy & DeShon, 2000; Schmidt, Viswesvaran, & Ones, 2000). In our study, we unconfounded the systematic effects that are commonly described in the literature as being critical to the measurement design of multisource job performance ratings (see Table 1). Our findings suggest that the measurement structure of multisource ratings is defined by (a) impressions related to different sources (e.g., managers, peers, etc) and (b) general performance. Specific dimensions had no practical relevance to this measurement structure.

Our findings suggest that the measurement design for job performance ratings should be reconsidered so as to formally acknowledge source-based and general perspectives on performance, rather than dimension-based perspectives. Taken to their logical conclusion, our results suggest that performance feedback should only be presented in terms of source-based scores and overall performance scores. Consideration should therefore be given to revising theory related to source-dependent dimensions judgments, for which we found no evidence. Dimensions present a conceptually intuitive framework for summarizing work-related information that might appeal to both employees and to organizational decision makers. Further investigation into how raters process information related to ratee performance is warranted with a view to testing whether dimension-related effects can be increased to a more meaningful magnitude. We nonetheless recommend that practitioners using multisource ratings data should not conclude that summary dimenson scores will provide a meaningful basis for performance evalation and promotion decisions.

References

Aguinis, H. (2019). *Performance management* (4th ed.). Chicago: Chicago Business Press.

Aguinis, H., Ji, Y. H., & Joo, H. (2018). Gender productivity gap among star performers in STEM and other scientific fields. *Journal of Applied Psychology, 103*, 1283-1306. doi: 10.1037/apl0000331

Bennett, W., Jr., Lance, C. E., & Woehr, D. J. (2006). *Performance measurement: Current perspectives and future challenges*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Bernardin, J. H., Konopaske, R., & Hagan, C. M. (2012). A comparison of adverse impact levels based on top-down, multisource, and assessment center data: Promoting diversity and reducing legal challenges. *Human Resource Management, 51*, 313-341. doi: 10.1002/hrm.21472

Borman, W. C. (1974). The rating of individuals in organizations: An alternate approach. *Organizational Behavior & Human Performance, 12*, 105-124. doi: 10.1016/0030-5073(74)90040-3

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10*, 99-109. doi: 10.1207/s15327043hup1002_3

Brennan, R. L. (2000). (Mis)conceptions about generalizability theory. *Educational Measurement: Issues and Practice, 19*, 5-10. doi: 10.1111/j.1745-3992.2000.tb00017.x

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

Church, A. H., & Bracken, D. W. (1997). Advancing the state of the art of 360-degree feedback: Guest editors' comments on the research and practice of multirater assessment methods. *Group and Organization Management, 22*, 149-161. doi: 10.1177/1059601197222002

Committee on Test Standards. (2011). Code of good practice for psychological testing. Leicester, UK: British Psychological Society.

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*, 1092-1122. doi: 10.1037/a0021212

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: 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*, 137-163. doi: 10.1111/j.2044-8317.1963.tb00206.x

DeNisi, A. S., & Murphy, K. R. (2017). Performance appraisal and performance management: 100 years of progress? *Journal of Applied Psychology, 102*, 421-433. doi: 10.1037%2Fapl0000085

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*, 515-533. doi: 10.1214/06-BA117A

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). New York: Chapman & Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457-472. doi: 10.1214/ss/1177011136

Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology, 83*, 960-968. doi: 10.1037/0021-9010.83.6.960

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*, 1-21. doi: 10.1111/j.1744-6570.2003.tb00141.x

Guenole, N., Cockerill, T., Chamorro-Premuzic, T., & Smillie, L. (2011). Evidence for the validity of 360 dimensions in the presence of rater-source factors. *Consulting Psychology Journal: Practice and Research, 63*, 203-218. doi: 10.1037/a0026537

Guion, R. (1965). *Personnel Testing*. New York: McGraw-Hill.

Hoffman, B. J., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology, 63*, 119-151. doi: 10.1111/j.1744-6570.2009.01164.x

Hoffman, B. J., Melchers, K. G., Blair, C. A., Kleinmann, M., & Ladd, R. T. (2011). Exercises and dimensions are the currency of assessment centers. *Personnel Psychology, 64*, 351-395. doi: 10.1111/j.1744-6570.2011.01213.x

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*, 1593-1623.

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*, 976-994. doi: 10.1037/apl0000102

Klimoski, R. J., & London, M. (1974). Role of the rater in performance appraisal. *Journal of Applied Psychology, 59*, 445-451. doi: 10.1037/h0037332

Knapp, D. J. (2006). The U.S. joint-service job performance measurement project. In W. Bennett, Jr., C. E. Lance & D. J. Woehr (Eds.), *Performance measurement: Current perspectives and future challenges.* (pp. 113-140). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Kraiger, K., & Teachout, M. S. (1990). Generalizability theory as construct-related evidence of the validity of job performance ratings. *Human Performance, 3*, 19. doi: 10.1207/s15327043hup0301_2

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15*, 722-752. doi: 10.1177/1094428112457829

Kuncel, N. R., & Sackett, P. R. (2012, April). *Resolving the assessment center construct validity problem.* Paper presented at the Society for Industrial and Organizational Psychology, San Diego, CA.

Lance, C. E. (2008). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 84-97. doi: 10.1111/j.1754-9434.2007.00017.x

Lance, C. E., Baxter, D., & Mahan, R. P. (2006). Evaluation of alternative perspectives on source effects in multisource performance measures. In W. Bennett Jr., C. E. Lance & D. J. Woehr (Eds.), *Performance Measurement: Current Perspectives and Future Challenges*. Malwah, NJ: Lawrence Erlbaum Associates.

Lance, C. E., Hoffman, B. J., Gentry, W. A., & Baranik, L. E. (2008). Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Human Resource Management Review, 18*, 223-232. doi: DOI 10.1016/j.hrmr.2008.03.002

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology, 89*, 377-385. doi: 10.1037/0021.9010.89.2.377

Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychological Methods, 7*, 228-244. doi: 10.1037%2F1082-989X.7.2.228

Lance, C. E., Teachout, M. S., & Donnelly, T. M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology, 77*, 437-452. doi: 10.1037/0021-9010.77.4.437

Le, H., Schmidt, F. L., & Putka, D. J. (2009). The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organizational Research Methods, 12*, 165-200. doi: Doi 10.1177/1094428107302900

Le, H., Schmidt., F. L., Harter, J. K., & Lauver, K. J. (2010). The problem of empirical redundancy of constructs in organizational research: An empirical investigation. *Organizational Behavior and Human Decision Processes, 112*, 112-125. doi: 10.1016/j.obhdp.2010.02.003

LeBreton, J. M., Scherer, K. T., & James, L. R. (2014). Corrections for criterion reliability in validity generalization: A false prophet in a land of suspended judgment. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 7*, 478-500. doi: 10.1111/iops.12184

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

Mount, M. K., Judge, T. A., Scullen, S. C., Sytsma, M. R., & Hezlett, S. A. (1998). Trait, rater, and level effects in 360-degree performance ratings. *Personnel Psychology, 51*, 557-575. doi: 10.1111/j.1744-6570.1998.tb00251.x

Murphy, K. R. (2008). Explaining the weak relationship between job performance and ratings of job performance. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 148-160. doi: 10.1111/j.1754-9434.2008.00030.x

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873-900. doi: 10.1111/j.1744-6570.2000.tb02421.x

Norman, W. T., & Goldberg, L. R. (1966). Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology, 4*, 681-691. doi: 10.1037/h0024002

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

Papaspiliopoulos, O., Roberts, G. O., & Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science, 22*, 59-73. doi: 10.1214/088342307000000014

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. *Organizational Behavior & Human decision Processes, 38*, 76-91. doi: 10.1016/0749-5978(86)90027-0

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*, 114-133. doi: 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). New York: Taylor & Francis.

Putka, D. J., Lance, C. E., Le, H., & McCloy, R. A. (2011). A cautionary note on modeling multitrait-multirater data arising from ill-structured measurement designs. *Organizational Research Methods, 14*, 503-529. doi: 10.1177/1094428110362107

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*, 959-981. doi: 10.1037/0021-9010.93.5.959

R Core Team. (2017). R: A language and environment for statistical computing (Version 3.4.1). Vienna, Austria: R Foundation for Statistical Computing.

Ree, M. J., Carretta, T. R., & Teachout, M. S. (2015). Pervasiveness of dominant general factors in organizational measurement. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8*, 409-427. doi: 10.1017/iop.2015.16

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rollan, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology, 88*, 1068-1081. doi: 10.1037/0021-9010.88.6.1068

Salgado, J. F., Anderson, N., & Tauriz, G. (2015). The validity of ipsative and quasi-ipsative forced-choice personality inventories for different occupational groups: A comprehensive meta-analysis. *Journal of Occupational and Organizational Psychology, 88*, 797-834. doi: 10.1111/joop.12098

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

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*, 956-970. doi: 10.1037//0021-9010.85.6.956

Searle, S. R., Casella, G., & McCulloch, C. E. (2006). *Variance components*. New York: Wiley.

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

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *The American Journal of Psychology, 18*, 161-169. doi: 10.2307/1412408

Stan Development Team. (2017). Stan: A C++ library for probability and sampling (Version 2.17.0).

Stan Development Team. (2018). RStan: the R interface to Stan (Version 2.17.3).

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*, 25-29. doi: 10.1037/h0071663

Toegel, G., & Conger, J. A. (2003). 360-degree assessment: Time for reinvention. *Academy of Management Learning and Education, 2*, 297-311. doi: 10.5465/AMLE.2003.10932156

Viswesvaran, C., Ones, D. S., Schmidt, F. L., Le, H., & Oh, I. (2014). Measurement error obfuscates scientific knowledge: Path to cumulative knowledge requires corrections for unreliability and psychometric meta-analyses. *Industrial & Organizational Psychology, 7*, 507-518. doi: 10.1111/iops.12186

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*, 108-131. doi: 10.1037/0021-9010.90.1.108

Williams, K. M., & Crafts, J. L. (1997). Inductive job analysis: The job/task inventory method. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp. 51-88). Palo Alto, CA: Davies-Black Publishing.

Zimmerman, R. D., Mount, M. K., & Goff, M., III.,. (2008). Multisource feedback and leaders' goal performance: Moderating effects of rating purpose, rater perspective, and performance dimension. *International Journal of Selection and Assessment, 16*, 121-133. doi: 10.1111/j.1468-2389.2008.00417.x

Zyphur, M. J. (2009). When mindsets collide: Switching analytical mindsets to advance organization science. *Academy of Management Review, 34*, 677-688. doi: 10.5465/AMR.2009.44885862

Zyphur, M. J., Oswald, F. L., & Rupp, D. E. (2015). Rendezvous overdue: Bayes analysis meets organizational research. *Journal of Management, 41*, 387-389. doi: 10.1177/0149206314549252

Table 1

Selected Cross-Study Comparison of Multiple Effects Estimated in Multisource Ratings

O'Neill et al.[a]		Hoffman et al.[b]		Scullen et al.(a)[c]		Scullen et al.(b)[d]		Greguras & Robie[e]		Kraiger & Teachout[f]	
Effect	%	Effect	%	Effect	%	Effect	%	Effect	%	Effect	%
p	23.00	p	3.50	p	13.00	p	14.00	p	24.03	p	12.14
pd	6.00	pd	7.10	pd	8.00	pd	11.00	-	-	-	-
-	-	-	-	-	-	-	-	-	-	s	2.43
-	-	ps	22.00	ps	9.00	ps	7.00	-	-	ps	33.74
-	-	-	-	-	-	-	-	-	-	pf	2.67
-	-	-	-	-	-	-	-	-	-	sf	0.12
-	-	-	-	-	-	-	-	-	-	psf	6.43
-	-	-	-	-	-	-	-	pi	4.10	p(i:f)	6.92
-	-	-	-	-	-	-	-	-	-	s(i:f)	0.36
r	23.00	-	-	-	-	-	-	-	-	-	-
rd	2.00	-	-	-	-	-	-	-	-	-	-
pr	25.00	pr	57.60	pr	62.00	pr	53.00	r,pr	28.23	-	-
prd,e	21.00	pds,e	14.80	pds,e	11.00	pds,e	18.00	ri,pri,r,e	43.64	ps(i:f),e	35.19

Note. Based on estimates presented in the original articles with reference to [a]O'Neill et al. (2015) Table 5; [b]B. J. Hoffman et al. (2010) Table 4, [c]Profiler data from Scullen et al. (2000) Table 3; [d]Management Skills Profile data from Scullen et al. Table 3; [e]between-participant-relevant effects and supervisor ratings only from Greguras and Robie (1998); [f]between-participant-relevant effects from Kraiger and Teachout (1990) Table 3. Note that Greguras and Robie also analyzed data separately from peers and subordinates, but did not model a source-related effect. The same measurement design as that in Greguras and Robie also appeared in Greguras et al. (2003). p = participant or ratee, d = dimension; r = rater; s = source; f = form or specific assessment content; i = item; e = residual error. The additional commas in the Greguras and Robie estimates indicate confounded effects (e.g., r,pr = r confounded with pr). Note that the Hoffman et al. and Scullen et al. effects add to just over 100%, presumably due to rounding error. Dashes account for cross-study differences in estimated effects and serve to align analogous or similar effects across studies. Note that several effects presented above are to be interpreted as analogues only of the original effects that were tested. E.g., Hoffman et al. did not specifically test a rater × participant interaction but they did test an analogue of this effect. The presentation above is intended as an aid to cross-study comparison. The presence of a colon indicates a level of nesting (e.g., i:f = items nested in forms).

Table 2

Sample 1: Variance Decomposition for Pre-Aggregated Ratings, Aggregation to Dimensions, and Dimensions across Raters in Sources

Effect	Pre-aggregation			Dimensions			Dimensions aggregated across raters in sources		
	VE	Total variance (%)	BP-variance (%)	Formula	VE	BP-variance (%)	Formula	VE	BP-variance (%)
BP-relevant									
σ^2_{p}	.0642	17.24	20.39	σ^2_{p}	.0642	27.23	σ^2_{p}	.0642	29.77
σ^2_{s}	.0649	17.42	20.60	σ^2_{s}	.0649	27.51	σ^2_{s}	.0649	30.07
$\sigma^2_{r:s}$	.0139	3.74	4.43	$\sigma^2_{r:s}$	.0139	5.91	$\sigma^2_{r:s}/n_{r:s}$	.0069	3.20
σ^2_{pd}	.0027	0.73	0.86	σ^2_{pd}	.0027	1.15	σ^2_{pd}	.0027	1.26
$\sigma^2_{pi:d}$	.0327	8.78	10.39	$\sigma^2_{pi:d}/n_{i:d}$	.0020	0.84	$\sigma^2_{pi:d}/n_{i:d}$	.0020	0.92
σ^2_{ps}	.0604	16.20	19.17	σ^2_{ps}	.0604	25.60	σ^2_{ps}	.0604	27.98
$\sigma^2_{pr:s}$	.0193	5.18	6.13	$\sigma^2_{pr:s}$	.0193	8.18	$\sigma^2_{pr:s}/n_{r:s}$	.0096	4.43
σ^2_{ds}	.0007	0.18	0.21	σ^2_{ds}	.0007	0.28	σ^2_{ds}	.0007	0.30
$\sigma^2_{dr:s}$	.0005	0.13	0.15	$\sigma^2_{dr:s}$	.0005	0.20	$\sigma^2_{dr:s}/n_{r:s}$	.0002	0.11
$\sigma^2_{is:d}$	.0056	1.51	1.79	$\sigma^2_{is:d}/n_{i:d}$	.0003	0.15	$\sigma^2_{is:d}/n_{i:d}$	.0003	0.16
$\sigma^2_{ir:ds}$	.0040	1.08	1.28	$\sigma^2_{ir:ds}/n_{i:d}$	.0002	0.10	$\sigma^2_{ir:ds}/n_{i:d}n_{r:s}$	.0001	0.06
$\sigma^2_{pdr:s}$	.0036	0.98	1.16	$\sigma^2_{pdr:s}$	.0036	1.55	$\sigma^2_{pdr:s}/n_{r:s}$	.0018	0.84
σ^2_{pds}	.0005	0.14	0.17	σ^2_{pds}	.0005	0.23	σ^2_{pds}	.0005	0.25
$\sigma^2_{pis:d}$	.0043	1.17	1.38	$\sigma^2_{pis:d}/n_{i:d}$	.0003	0.11	$\sigma^2_{pis:d}/n_{i:d}$	.0003	0.12
$\sigma^2_{pir:ds,e}$	.0375	10.07	11.91	$\sigma^2_{pir:ds,e}/n_{i:d}$	.0023	0.97	$\sigma^2_{pir:ds,e}/n_{i:d}n_{r:s}$	.0011	0.52
BP-non-relevant									
σ^2_{d}	.0104	2.79	–	σ^2_{d}	–	–	σ^2_{d}	–	–
$\sigma^2_{i:d}$	.0472	12.67	–	$\sigma^2_{i:d}$	–	–	$\sigma^2_{i:d}$	–	–

Note. VE = variance estimate, BP = between-participant, p = participants (i.e., ratees), s = sources, r = raters, d = dimensions, i = rating items. All effects involving raters were corrected with the *q*-multiplier for ill-structured designs. The presence of a colon indicates a level of nesting (e.g., i:d = rating items nested in dimensions). The denominator r:s reflects the grand mean of the ratee means of the rater source frequencies.

Table 3

Sample 1: Generalization for Pre-Aggregated Ratings, Aggregation to Dimensions, and to Dimensions and Raters in Sources

Aggregation level/	Classification of variance components		
Intended generalization	Universe score	Error	$E\rho^2$
Pre-aggregation			
s, r	p, pd, pi:d	s, r:s, ps, pr:s, ds, dr:s, is:d, ir:ds, pdr:s, pds, pis:d, pir:ds	.33
r	p, s, pd, pi:d, ps, ds, is:d, pds, pis:d	r:s, pr:s, dr:s, ir:ds, pdr:s, pir:ds	.74
i, r	p, s, pd, ps, ds, pds	r:s, pi:d, pr:s, dr:s, is:d, ir:ds, pdr:s, pis:d, pir:ds	.60
i, s, r	p, pd	s, r:s, pi:d, ps, pr:s, ds, dr:s, is:d, ir:ds, pdr:s, pds, pis:d, pir:ds	.22
Dimensions			
s, r	p, pd, pi:d/$n_{i:d}$	s, r:s, ps, pr:s, ds, dr:s, is:d/$n_{i:d}$, ir:ds/$n_{i:d}$, pdr:s, pds, pis:d/$n_{i:d}$, pir:ds/$n_{i:d}$	.31
r	p, s, pd, pi:d/$n_{i:d}$, ps, ds, is:d/$n_{i:d}$, pds, pis:d/$n_{i:d}$	r:s, pr:s, dr:s, ir:ds/$n_{i:d}$, pdr:s, pir:ds/$n_{i:d}$	.82
i, r	p, s, pd, ps, ds, pds	r:s, pi:d/$n_{i:d}$, pr:s, dr:s, is:d/$n_{i:d}$, ir:ds/$n_{i:d}$, pdr:s, pis:d/$n_{i:d}$, pir:ds/$n_{i:d}$	.81
i, s, r	p, pd	s, r:s, pi:d/$n_{i:d}$, ps, pr:s, ds, dr:s, is:d/$n_{i:d}$, ir:ds/$n_{i:d}$, pdr:s, pds, pis:d/$n_{i:d}$, pir:ds/$n_{i:d}$	.30
Dimensions, raters in sources			
s, r	p, pd, pi:d/$n_{i:d}$	s, r:s/$n_{r:s}$, ps, pr:s/$n_{r:s}$, ds, dr:s/$n_{r:s}$, is:d/$n_{i:d}$, ir:ds/$n_{i:d}n_{r:s}$, pdr:s/$n_{r:s}$, pds, pis:d/$n_{i:d}$, pir:ds/$n_{i:d}n_{r:s}$	.34
r	p, s, pd, pi:d/$n_{i:d}$, ps, ds, is:d/$n_{i:d}$, pds, pis:d/$n_{i:d}$	r:s, pr:s/$n_{r:s}$, dr:s/$n_{r:s}$, ir:ds/$n_{i:d}n_{r:s}$, pdr:s/$n_{r:s}$, pir:ds/$n_{i:d}n_{r:s}$	.90
i, r	p, s, pd, ps, ds, pds	r:s/$n_{r:s}$, pi:d/$n_{i:d}$, pr:s/$n_{r:s}$, dr:s/$n_{r:s}$, is:d/$n_{i:d}$, ir:ds/$n_{i:d}n_{r:s}$, pdr:s/$n_{r:s}$, pis:d/$n_{i:d}$, pir:ds/$n_{i:d}n_{r:s}$	.89
i, s, r	p, pd	s, r:s/$n_{r:s}$, pi:d/$n_{i:d}$, ps, pr:s/$n_{r:s}$, ds, dr:s/$n_{r:s}$, is:d/$n_{i:d}$, ir:ds/$n_{i:d}n_{r:s}$, pdr:s/$n_{r:s}$, pds, pis:d/$n_{i:d}$, pir:ds/$n_{i:d}n_{r:s}$	.33

Note. p = participants (i.e., ratees), s = sources, r = raters, d = dimensions, i = rating items, $E\rho^2$ = expected reliability. To interpret each row: for example, dimensions, s,r means that the specified row relates to items aggregated to dimension scores and that the intent is to generalize across both sources and raters. The presence of a colon indicates a level of nesting (e.g., i:d = rating items nested in dimensions). The denominator r:s reflects the grand mean of the ratee means of the rater source frequencies. All effects involving raters were corrected with the q-multiplier for ill-structured designs. Note that in this design, generalization to items and sources cannot be statistically differentiated from generalization to items, sources, and raters, hence there is no separate entry for i, s. All coefficients presented here were generated from posterior distributions.

Table 4

Sample 2: Variance Decomposition for Pre-Aggregated Ratings, and Aggregation to Dimensions, and Dimensions across Raters in Sources

Effect	Pre-aggregation			Dimensions			Dimensions aggregated across raters in sources		
	VE	Total variance (%)	BP-variance (%)	Formula	VE	BP variance (%)	Formula	VE	BP variance (%)
BP-relevant									
σ^2_p	.0383	12.08	19.59	σ^2_p	.0383	38.23	σ^2_p	.0383	39.22
σ^2_s	.0117	3.69	5.99	σ^2_s	.0117	11.68	σ^2_s	.0117	11.98
σ^2_{ps}	.0192	6.07	9.85	σ^2_{ps}	.0192	19.22	σ^2_{ps}	.0192	19.71
σ^2_{pc}	.0098	3.10	5.03	σ^2_{pc}	.0098	9.81	σ^2_{pc}	.0098	10.07
$\sigma^2_{pd:c}$	.0028	0.90	1.45	$\sigma^2_{pd:c}$	.0028	2.83	$\sigma^2_{pd:c}$	.0028	2.91
σ^2_{psc}	.0013	0.42	0.68	σ^2_{psc}	.0013	1.33	σ^2_{psc}	.0013	1.37
$\sigma^2_{psd:c}$	.0011	0.34	0.55	$\sigma^2_{psd:c}$	.0011	1.07	$\sigma^2_{psd:c}$	.0011	1.10
$\sigma^2_{si:d:c}$	.0062	1.96	3.18	$\sigma^2_{si:d:c}/\mathrm{n}_{\mathrm{i:d}}$	.0006	0.62	$\sigma^2_{si:d:c}/\mathrm{n}_{\mathrm{i:d}}$	.0006	0.63
$\sigma^2_{pi:d:c}$	.0661	20.87	33.85	$\sigma^2_{pi:d:c}/\mathrm{n}_{\mathrm{i:d}}$	.0066	6.58	$\sigma^2_{pi:d:c}/\mathrm{n}_{\mathrm{i:d}}$	.0066	6.75
$\sigma^2_{psi:d:c}$	.0171	5.41	8.77	$\sigma^2_{psi:d:c}/\mathrm{n}_{\mathrm{i:d}}$	.0017	1.71	$\sigma^2_{psi:d:c}/\mathrm{n}_{\mathrm{i:d}}$	.0017	1.75
σ^2_{sc}	.0002	0.05	0.08	σ^2_{sc}	.0002	0.17	σ^2_{sc}	<.0001	0.01
$\sigma^2_{sd:c}$	.0004	0.13	0.20	$\sigma^2_{sd:c}$	.0004	0.40	$\sigma^2_{sd:c}$	.0003	0.26
$\sigma^2_{r:ps}$	.0042	1.32	2.15	$\sigma^2_{r:ps}$	.0042	4.19	$\sigma^2_{r:ps}/\mathrm{n}_{\mathrm{r:s}}$	.0027	2.79
$\sigma^2_{rc:ps}$	.0004	0.14	0.22	$\sigma^2_{rc:ps}$	.0004	0.43	$\sigma^2_{rc:ps}/\mathrm{n}_{\mathrm{r:s}}$	.0003	0.28
$\sigma^2_{rd:pcs}$	.0001	0.04	0.06	$\sigma^2_{rd:pcs}$	.0001	0.12	$\sigma^2_{rd:pcs}/\mathrm{n}_{\mathrm{r:s}}$	.0001	0.08
$\sigma^2_{ri:psd:c,e}$	.0163	5.14	8.33	$\sigma^2_{ri:psd:c,e}/\mathrm{n}_{\mathrm{i:d}}$	.0016	1.62	$\sigma^2_{ri:psd:c,e}/\mathrm{n}_{\mathrm{i:d}}\mathrm{n}_{\mathrm{r:s}}$	.0011	1.08
BP-non-relevant									
σ^2_c	.0106	3.34	–	σ^2_c	–	–	σ^2_c	–	–
$\sigma^2_{d:c}$	.0070	2.20	–	$\sigma^2_{d:c}$	–	–	$\sigma^2_{d:c}$	–	–
$\sigma^2_{i:d:c}$	.1039	32.80	–	$\sigma^2_{i:d:c}$	–	–	$\sigma^2_{i:d:c}$	–	–

Note. VE = variance estimate, BP = between-participant, p = participants (i.e., ratees), s = sources, r = raters, d = dimensions, i = rating items, c = competency categories. All effects involving raters were corrected with the q-multiplier for ill-structured designs. The presence of a colon indicates a level of nesting (e.g., i:d = rating items nested in dimensions). The denominator r:s reflects the grand mean of the ratee means of the rater source frequencies.

Table 5

Sample 2: Generalization for Pre-Aggregated Ratings, Aggregation to Dimensions, and Dimensions across Raters in Sources

Aggregation/G	Classification of variance components		
	Universe score	Error	$E\rho^2$
Pre-aggregation			
s, r	p, pc, pd:c, pi:d:c	s, ps, psc, psd:c, si:d:c, psi:d:c, sc, sd:c, r:ps, rc:ps, rd:pcs, ri:psd:c,e	.60
r	p, s, ps, pc, pd:c, psc, psd:c, si:d:c, pi:d:c, psi:d:c, sc, sd:c	r:ps, rc:ps, rd:pcs, ri:psd:c,e	.89
i, r	p, s, ps, pc, pd:c, psc, psd:c, sc, sd:c	si:d:c, pi:d:c, psi:d:c, r:ps, rc:ps, rd:pcs, ri:psd:c,e	.43
i, s, r	p, pc, pd:c	s, ps, psc, psd:c, si:d:c, pi:d:c, psi:d:c, sc, sd:c, r:ps, rc:ps, rd:pcs, ri:psd:c,e	.26
Dimensions			
s, r	p, pc, pd:c, pi:d:c/$n_{i:d}$	s, ps, psc, psd:c, si:d:c/$n_{i:d}$, psi:d:c/$n_{i:d}$, sc, sd:c, r:ps, rc:ps, rd:pcs, ri:psd:c,e/$n_{i:d}$	.59
r	p, s, ps, pc, pd:c, psc, psd:c, si:d:c/$n_{i:d}$, pi:d:c/$n_{i:d}$, psi:d:c/$n_{i:d}$, sc, sd:c	r:ps, rc:ps, rd:pcs, ri:psd:c,e/$n_{i:d}$	.93
i, r	p, s, ps, pc, pd:c, psc, psd:c, sc, sd:c	si:d:c/$n_{i:d}$, pi:d:c/$n_{i:d}$, psi:d:c/$n_{i:d}$, r:ps, rc:ps, rd:pcs, ri:psd:c,e/$n_{i:d}$	.84
i, s, r	p, pc, pd:c	s, ps, psc, psd:c, si:d:c/$n_{i:d}$, pi:d:c/$n_{i:d}$, psi:d:c/$n_{i:d}$, sc, sd:c, r:ps, rc:ps, rd:pcs, ri:psd:c,e/$n_{i:d}$	.52
Dimensions, raters in sources			
s, r	p, pc, pd:c, pi:d:c/$n_{i:d}$	s, ps, psc, psd:c, si:d:c/$n_{i:d}$, psi:d:c/$n_{i:d}$, sc, sd:c, r:ps/$n_{r:s}$, rc:ps/$n_{r:s}$, rd:pcs/$n_{r:s}$, ri:psd:c,e/$n_{i:d}n_{r:s}$	.60
r	p, s, ps, pc, pd:c, psc, psd:c, si:d:c/$n_{i:d}$, pi:d:c/$n_{i:d}$, psi:d:c/$n_{i:d}$, sc, sd:c	r:ps/$n_{r:s}$, rc:ps/$n_{r:s}$, rd:pcs/$n_{r:s}$, ri:psd:c,e/$n_{i:d}n_{r:s}$	.96
i, r	p, s, ps, pc, pd:c, psc, psd:c, sc, sd:c	si:d:c/$n_{i:d}$, pi:d:c/$n_{i:d}$, psi:d:c/$n_{i:d}$, r:ps/$n_{r:s}$, rc:ps/$n_{r:s}$, rd:pcs/$n_{r:s}$, ri:psd:c,e/$n_{i:d}n_{r:s}$	.86
i, s, r	p, pc, pd:c	s, ps, psc, psd:c, si:d:c/$n_{i:d}$, pi:d:c/$n_{i:d}$, psi:d:c/$n_{i:d}$, sc, sd:c, r:ps/$n_{r:s}$, rc:ps/$n_{r:s}$, rd:pcs/$n_{r:s}$, ri:psd:c,e/$n_{i:d}n_{r:s}$	.53

Note. p = participants (i.e., ratees), s = sources, r = raters, d = dimensions, i = rating items, c = competency categories, $E\rho^2$ = expected reliability, Aggregation = aggregation level, G = intended generalization. To interpret each row: for example, dimensions, s,r means that the specified row relates to items aggregated to dimension scores and that the intent is to generalize across both sources and raters. The presence of a colon indicates a level of nesting (e.g., i:d = rating items nested in dimensions). The denominator r:s reflects the grand mean of the ratee means of the rater source frequencies. Note that in this design, generalization to rating items and sources cannot be statistically differentiated from generalization to rating items, sources, and raters, hence there is no separate entry for i, s. All coefficients presented here were generated from posterior distributions.

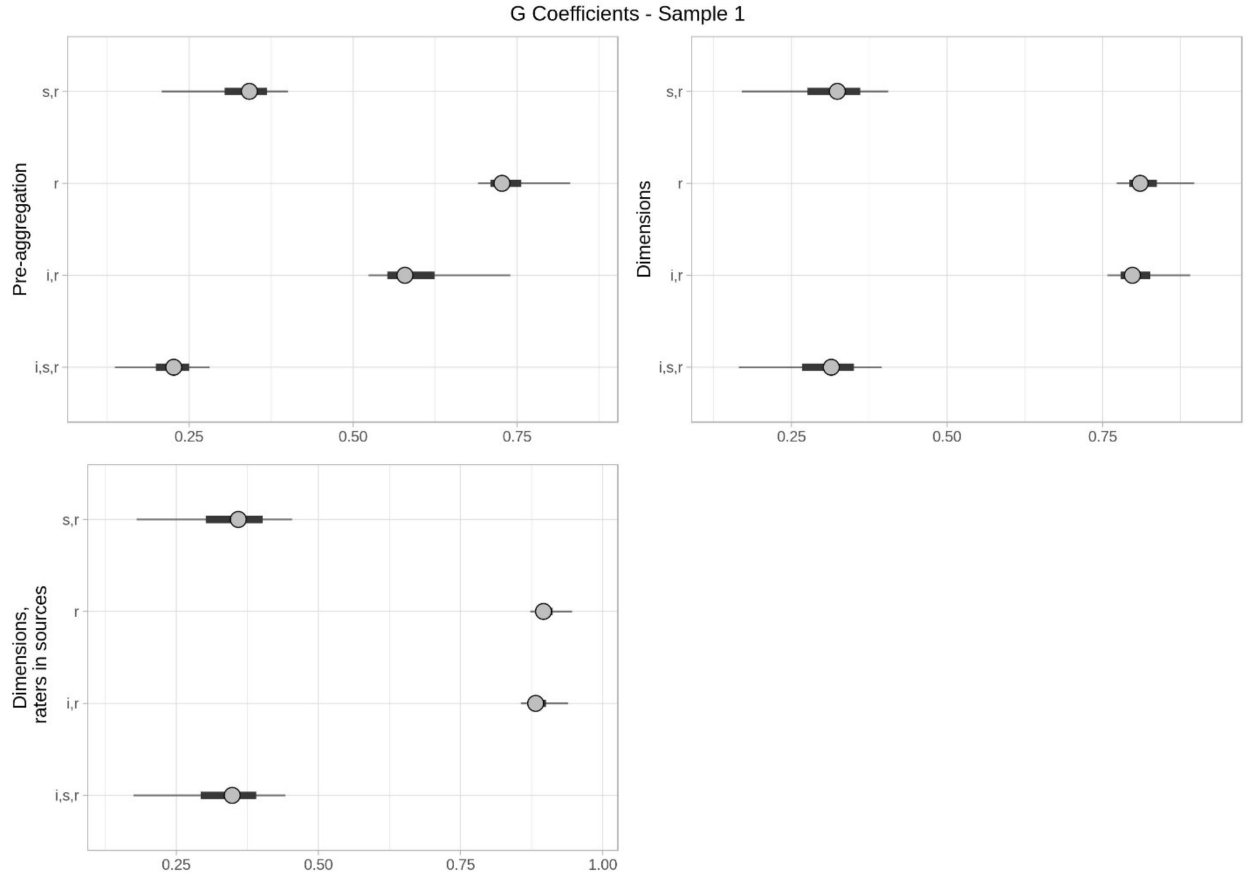

Figure 1. Generalizability (G) coefficients and credible intervals for different types of aggregation level and generalization across combinations of sources (s), raters (r), and items (i) from Sample 1.

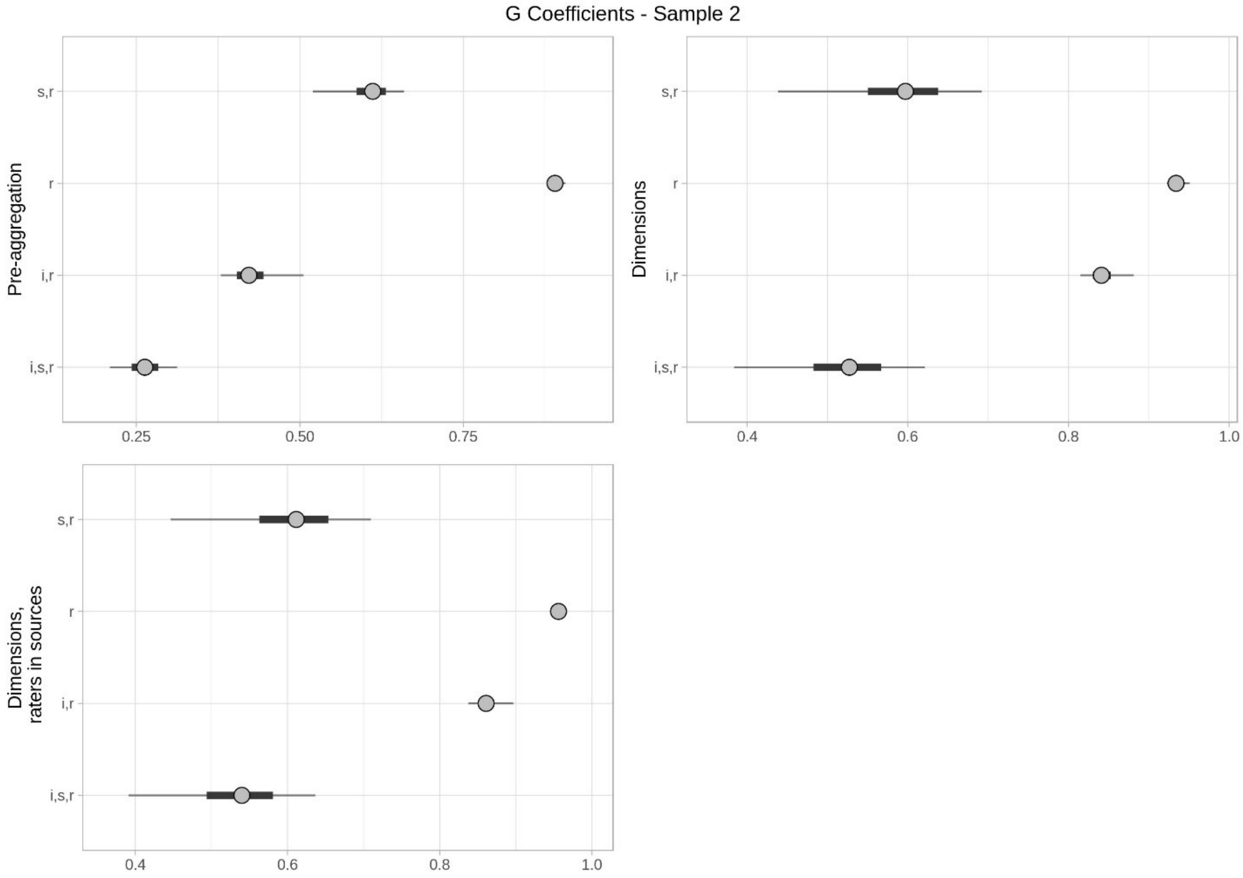

Figure 2. Generalizability (G) coefficients and credible intervals for different types of aggregation level and generalization across combinations of sources (s), raters (r), and items (i) from Sample 2.

Appendix

Table A1

Dimension Definitions by Sample

Sample 1	Dimension	Definition
	Teamwork	Able to work collaboratively towards goal achievement
	Organizational Citizenship	Exceeding expectations regarding work output
	Results-focused	Centered on achieving results for the organization
	Motivation	Sustaining a high work-oriented energy level
Sample 2	Broad Dimension/Dimension	Definition
	Goal setting	
	Delegating	Ability to allocate tasks to suitable individuals
	Independence	Autonomy and forthrightness in expressing views
	Managing change	Accountability for and successfully implementing change
	Persuasive communication	Able to express self, influence others, and negotiate
	Project management	Driven to achieve project objectives
	Results orientation	Task-oriented and focused on job completion
	Organizational	
	Attention to detail	Taking care with tasks and complying with procedure
	Commitment	Identifying with organizational objectives and values
	Information management	Using research and facts to help guide decision making
	Planning and organizing	Devising effective processes and procedures
	Interpersonal	
	Communication skills	Ability to engage with and appreciate needs of others
	Customer focus	Capacity to engage effectively with customers
	Developing others	Inspiring confidence and growth in others
	Interpersonal skills	Sensitivity and appreciation of emotional needs
	People management	Ability to manage others effectively
	Team orientation	Collaborating effectively and positively with others
	Enterprise	
	Leadership potential	Tolerance, determination, motivating, and inspiring vision
	Motivation	Drive and aspiration to succeed
	Resilience	Coping with stress and remaining calm under pressure
	Risk-taking	Spontaneity, excitement-seeking, challenging convention
	Self-confidence	Being sure of oneself, optimistic, and up-beat
	Strategy	
	Analytic	Systematic and considered problem solving style
	Creative	Curious, divergent thinking, bringing new perspectives
	Decision making	Able to balance caution with appropriate risk-taking
	Flexibility	Ability to cope with the unexpected
	Problem solving	Delivering effective, practical solutions to problems
	Strategic awareness	Able to objectively appraise events for strategic advantage

Note. Summary versions of the original definitions appear above. The broad dimensions in Sample 2 are defined by the specific dimensions nested in each.