Routledge
Taylor & Francis Group

Human Performance

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hhup20

Uncertainty about Rater Variance and Small Dimension Effects Impact Reliability in Supervisor Ratings

Duncan Jackson, George Michaelides, Christopher Dewberry, Amanda Jones, Simon Toms, Ben Schwencke & Wei-Ning Yang

To cite this article: Duncan Jackson, George Michaelides, Christopher Dewberry, Amanda Jones, Simon Toms, Ben Schwencke & Wei-Ning Yang (2022): Uncertainty about Rater Variance and Small Dimension Effects Impact Reliability in Supervisor Ratings, Human Performance, DOI: 10.1080/08959285.2022.2111433

To link to this article: https://doi.org/10.1080/08959285.2022.2111433

Published online: 19 Aug 2022.

HUMAN PERFORMANCE
https://doi.org/10.1080/08959285.2022.2111433

Routledge
Taylor & Francis Group

Uncertainty about Rater Variance and Small Dimension Effects Impact Reliability in Supervisor Ratings

Duncan Jackson[a], George Michaelides[b], Christopher Dewberry[c], Amanda Jones[a], Simon Toms[d], Ben Schwencke[e], and Wei-Ning Yang[a]

[a]King's College London, King's Business School; [b]University of East Anglia, Norwich Business School; [c]Not Applicable, Independent Scholar; [d]Psychological Consultancy Ltd, Consulting, Tunbridge Wells; [e]Test Partnership Ltd, Consulting

ABSTRACT

We modeled the effects commonly described as defining the measurement structure of supervisor performance ratings. In doing so, we contribute to different theoretical perspectives, including components of the multifactor and mediated models of performance ratings. Across two reanalyzed samples (Sample 1, N_{ratees} = 392, N_{raters} = 244; Sample 2, N_{ratees} = 342, N_{raters} = 397), we found a structure primarily reflective of general (>27% of variance explained) and rater-related (>49%) effects, with relatively small performance dimension effects (between 1% and 11%). We drew on findings from the assessment center literature to approximate the proportion of rater variance that might theoretically contribute to reliability in performance ratings. We found that even moderate contributions of rater-related variance to reliability resulted in a sizable impact on reliability estimates, drawing them closer to accepted criteria.

Performance ratings hold a central role in applied psychology and human resource management as a developmental aid, an indicator of individual performance, and as a criterion in validation studies (Aguinis, 2019; DeNisi & Murphy, 2017; Murphy & Cleveland, 1995; O'Neill, McLarnon, & Carswell, 2015). The performance of employees is often evaluated by their supervisors. A long-term concern related to supervisor ratings is the substantial accumulated evidence that they lack adequate reliability (Murphy, 2008; Thorndike, 1920). Researchers in the field estimate the interrater reliability coefficient of supervisor ratings as only around .52 (Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990; Schmidt, Viswesvaran, & Ones, 2000; Viswesvaran, Ones, & Schmidt, 1996); a figure well below levels typically regarded as acceptable in the psychometric literature (e.g., Lance, Butts, & Michels, 2006; LeBreton, Scherer, & James, 2014).

The .52 estimate for the reliability of performance ratings assumes that all rater-related variance in the evaluation of individual performance contributes to unreliability. This assumption has been challenged by researchers (Murphy & DeShon, 2000a, 2000b; Putka, Hoffman, & Carter, 2014), who argue that between-rater differences may arise from raters being exposed to, or focusing on, different aspects of each ratee's performance. This position implies that variability between raters might not only reflect variance that contributes to unreliability, but also reliable information of value to the evaluation of ratees.

To progress an understanding of these issues, it would be of assistance to establish the underlying measurement structure of performance ratings and to identify the degree of variance directly associated rater-related effects. Once rater-related effects have been modeled and their magnitude

CONTACT Duncan Jackson duncan.jackson@kcl.ac.uk King's Business School, King's College London, Bush House 30 Aldwych, London WC2B 4BG, United Kingdom of Great Britain and Northern Ireland

estimated, it is then necessary to estimate the proportion of between-rater variance associated with (a) raters reliably focusing on different aspects of a ratee performance, and (b) rater variance contributing to unreliability.

The present study contributes to the literature in four principal ways. First, we model the elements often described as comprising the measurement structure of supervisory ratings. In doing so, we provide a novel contribution to theory concerning the measurement structure of performance ratings. Second, we offer practice-relevant guidance about how each component of the measurement design (e.g., items, dimensions etc.) contributes to variance in ratings. Third, we approximate the proportion of rater variance that might represent reliable rater sensitivity to ratee performance variability versus rater variance that might contribute to unreliability. Fourth, the multifaceted approach we follow contributes to knowledge about the multifactor and, in part, to the mediated models for performance ratings summarized in Murphy (2008). The multifactor model suggests that performance ratings are affected by multiple, nonperformance factors. The mediated model suggests that performance ratings are further affected by rater goals and intentions.

The prevailing interrater reliability perspective

Two decades ago, Schmidt et al. (2000, p. 909) stated that interrater reliability "is the *only* appropriate reliability coefficient" for the purposes of correcting an observed validity correlation. Schmidt et al. based this assertion on the premise that interrater reliability corrects for four sources of measurement error, namely rater leniency, halo effects, random response error, and transient error. Leniency reflects the degree with which raters practice undue clemency or severity in their ratings. Halo effects refer to a general impression formed about an assessee, albeit positive or negative, or refers to a tendency to evaluate conceptually related dimensions similarly across ratees. Random response error refers to residual, non-systematic variability. By transient error, Schmidt et al. refer to within-assessor variance that might arise from the same assessor rating across different occasions of measurement.[1] Transient error is distinguished as a temporal effect, such that a rater's "mood, feeling, mental efficiency, or mental state" might vary on different occasions (Schmidt et al., 2000, p. 907).

Following this perspective, the Viswesvaran et al. (1996) meta-analysis provides a classic standpoint on the reliability of overall supervisor ratings (Putka & Hoffman, 2014). Viswesvaran et al. investigated interrater and intrarater reliability relating to 10 separate job performance dimensions and overall job performance. They estimated the interrater reliability of overall job performance, based on 40 studies, to be .52. This statistic has become the focus of much subsequent research and discussion (e.g., summarized in Murphy, 2003, 2008). It has moreover become the common estimate of choice in the correction of criterion-related validity coefficients involving supervisor ratings in both validity generalization and individual validation studies (Putka & Hoffman, 2014). This is particularly the case when study-specific reliability estimates are unavailable (LeBreton et al., 2014).

Concerns about the reliability of performance ratings

The low interrater reliability coefficient of .52 for supervisor ratings found by Viswesvaran et al. (1996) and other scholars was not entirely unexpected. Concerns about the measurement characteristics of performance ratings have been raised for over a century. Commenting on an organizational rating system involving multiple performance dimensions, Thorndike (1920, p. 25) stated that raters were "unable to analyze out these different aspects of the person's nature and achievement and rate each in independence of the others." Thorndike hypothesized that raters only assess in terms of general or perhaps halo-type judgments and cannot differentiate between specific performance attributes.

[1]Usually, but not exclusively, the source of variance related to occasions is estimated in classical psychometrics with test-retest reliability estimates.

In more recent literature, and touching on a related criticism, Murphy (2008) discussed the tenuous correspondence between performance and performance ratings, suggesting that such ratings fail to fulfil their intended purpose. Even within the last decade, renewed interest in and debate surrounding the status of performance ratings emerged with a focal article by LeBreton et al. (2014). LeBreton et al. raised the question: "Why are performance ratings allowed to survive in spite of what most would agree is abjectly problematic measurement?" (p. 482). The authors described performance ratings as "fundamentally flawed" and in which "~50% of the observed variance is measurement error" (p. 482).

Generalizability (G) theory and reliability estimation

The Schmidt et al. (2000) contention that interrater reliability is the only estimate relevant to corrections to unreliability in performance ratings has not gone without challenge. In particular, Murphy and DeShon (2000a) suggested the application of generalizability theory (G theory) to estimate reliability in this context. Unlike the process by which interrater reliability is traditionally estimated, G theory can be used to simultaneously model multiple effects relevant to the measurement structure of performance ratings. This allows for direct, statistically partialled comparisons between key variance components, including those relating to general, behavioral rating, dimension, and rater effects. G theory permits researchers to classify sources of rater-related variance as a contribution to unreliability or, alternatively, as reliable, systematic effects reflective of the rater's perspective on a given ratee. Unlike classical approaches to psychometrics, it thus facilitates researcher decisions on how to define multiple sources of universe (akin to true) score in contrast to multiple potential sources of unreliability and other, uncategorized sources variance[2] (Brennan, 2001). Moreover, G theory can be used to provide a detailed evaluation of how multiple measurement-design-relevant effects uniquely contribute to reliable and which contribute to unreliable variance,[3] and thus has the potential to inform theory on the structure of performance ratings.

With a G theory approach, once a complete set of effects relevant to a measurement design is estimated, it is possible to approximate the consequence of aggregating ratings into different types of summary score. Aggregation can have the effect of changing the proportion of variance associated with specific effects in a measurement structure (Putka & Hoffman, 2013, 2014). Sets of rating items might be aggregated to form dimension scores, dimension scores could then be aggregated across different raters, or aggregation could occur across all rating items, dimensions, and raters to arrive at overall scores. All three of these approaches to aggregation could result in different reliability outcomes, as has been suggested in other research contexts (Jackson, Michaelides, Dewberry, & Kim, 2016; Putka & Hoffman, 2013).

Extant G theory analyses of performance ratings

The measurement design for performance ratings is typically described as involving raters evaluating assessees on rating items nested in each of several performance dimensions (Bennett, Lance, & Woehr, 2006; Murphy & Cleveland, 1995; Murphy & DeShon, 2000a; O'Neill et al., 2015). We were unable to find an analysis that partialled effects discussed in the literature as being primarily relevant to this design (i.e., inclusive of raters, ratees, items, and performance dimensions). In the empirical studies we reviewed that investigated multiple effects, rater-related variance was always treated as contributing to unreliable variance (see Putka & Hoffman, 2013 for further discussion on this issue). The idea that at least some portion of rater variance might contribute to universe score (see Murphy & DeShon, 2000a, 2000b; Putka et al., 2014) does not appear to have been investigated empirically in this context.

[2]Uncategorized sources of variance are those that are neither relevant to universe score nor to unreliability, irrespective of the measurement intentions of the researcher. For example, when comparing across ratees, the main effect for items has no bearing on the rank ordering of ratees and is therefore neither relevant to universe score nor unreliable variance.

[3]We adopt the terms "reliable" (or universe score) and "unreliable" variance from Putka and Hoffman (2013).

An element of the performance ratings measurement design intended to directly summarize performance is that concerning performance dimensions (or "competencies"). Supervisors often evaluate ratees on dimensions such as *teamwork* and *communication skills* (e.g., Bartram, 2005; Kurz & Bartram, 2002). Greguras and Robie (1998) presented a G theory model that addressed several important sources of variance relevant to performance ratings. While the authors modeled item effects, they did not model performance dimensions. Moreover, raters were confounded with ratees in their design. Although central to their measurement design, the structure of performance dimensions (e.g., teamwork ability, customer focus) has generally been underexplored in the context of performance ratings. However, in many real-world measurement designs, performance dimensions play a central role, even in the estimation of overall scores. This is true of supervisory job performance ratings (Bartram, 2005), assessment center (AC) ratings (Putka & Hoffman, 2013), and situational judgment tests (Christian, Edwards, & Bradley, 2010). Dimensions are of theoretical importance because they supposedly define meaningful subcomponents of the performance construct domain (e.g., Arthur & Villado, 2008; Bartram, 2005; Borman & Brush, 1993).

O'Neill et al. (2015) is a rare example of the modeling of dimension effects, along with assessee and rater effects, for supervisor ratings. They found small dimension effects (around 6% of variance explained). However, in their study, item-related effects were not modeled. Item effects might play a key role in performance ratings, particularly given their involvement in aggregation relating to summative dimension scores.

Formulae for estimating the effects of aggregation are available in the G theory literature (e.g., Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). Putka and Hoffman (2014) adapted such formulae in their reanalysis of data from Greguras and Robie (1998). While insightful, the conclusions that could be drawn from this analysis were limited by the fact that, in the original study, assessees were confounded with raters and no performance dimensions were defined. In contrast, the effects in the O'Neill et al. (2015) study, while including dimensions, neither acknowledged different types of aggregation nor, as mentioned above, effects relating to rating items. Jackson, Michaelides, Dewberry, Schwencke, and Toms (2020) considered G theory formulae for aggregation as it relates to multisource ratings. However, we were unable to find a study that explored aggregation pertaining to an unconfounded measurement design specifically for supervisor performance ratings.

Theoretical models for performance ratings

G theory involves partitioning multiple measurement design effects that are potentially relevant or irrelevant to the performance construct (Cronbach et al., 1972). This approach facilitates exploration of the *multifactor* and *mediated* theoretical perspectives that have been proposed for performance and performance ratings (discussed below). Murphy (2008) summarized 3 theoretical models that describe the relationship between the performance construct and performance ratings. First, the *one-factor* model suggests a direct relationship between performance and ratings of performance. However, this relationship is subject to measurement error, which, if removed, allows for a direct representation of performance via ratings. The one-factor model assumes that performance ratings can be decomposed into true score + error, and corrections for the latter enable an estimation of performance (e.g., as in the corrections for attenuation in Viswesvaran et al., 1996).

Second, *multifactor* models aim to delineate a multiplicity of effects that might influence performance ratings, including performance, raters, items, job characteristics, and cognitive processes (Landy & Farr, 1980). Murphy suggests that a useful contribution from this work is that it highlights the impact not only of job performance, but also various other systematic factors on ratings. Multifactor models have the potential to help explain the array of nonperformance-related factors that may have a bearing on ratings. This idea has implications for shared rater variance relating to

estimates of general ratee performance. Such estimates might not only indicate ratee performance, but other, nonperformance characteristics, including individual (e.g., rater recall of events) and system characteristics (e.g., use of rating scales, see Murphy, Cleveland, & Hanscom, 2019).

Third, Murphy describes *mediated* models. These expand on multifactor models by suggesting that distortions (e.g., concerning organizational politics and individual rater goals) can influence the link between performance and ratings. The idea here is that the multiple influences identified in multifactor models are mediated through rater goals and intentions, which are, in turn, reflected in ratings. However, only one of the many factors involved in this evaluation and perceptual process is the performance of ratees.

To date, studies of the reliability of supervisor ratings have typically been conducted within the framework of classical test theory and thus align closely with the one-factor model described above. The classical approach typically involves correlating ratings provided by large numbers of supervisor pairs, where each pair evaluates the performance of a specific ratee. This provides a suitable, unbiased inter-rater reliability estimate of the ratings of the overall performance of ratees. However, it yields an incomplete perspective on performance ratings (Murphy & DeShon, 2000a, 2000b; Putka & Hoffman, 2014). This is because the design of supervisory ratings involves measurement elements that are ignored by the approach to reliability assumed in the one-factor model. As suggested in multifactor and mediated models, many of these elements are likely nonperformance effects that have a bearing on performance ratings.

Measurement design elements related and unrelated to performance

Rater-related effects have presented a topic of much debate. The most common approach taken in the literature is to treat all rater-related variance as a contribution to unreliable variance. The idea that rater variance contributes to unreliability is implied in the common estimation of and correction for interrater reliability estimates (Schmidt & Hunter, 1996; Viswesvaran et al., 1996). However, this is not the only perspective on the topic. Murphy and DeShon (2000a) suggest that there is "no clear justification" (p. 877) for defaulting to an "unreliable" classification for rater-related variance. They submit that rater perspectives on a given ratee might vary meaningfully because of the rater's position, their relationship with the ratee, and political motivations. Thus, rater variance could, in part, reflect different contextual perspectives on employee performance (Putka et al., 2014). To illustrate, one supervisor might have more experience with an employee in the context of client engagement. A different supervisor might have more experience with the same employee in the context of logistics management. These are different environments across which employee performance might meaningfully vary. Experience is only one example of the more general issue of variability in performance output that could be affected by any number of effects (e.g., stimuli, mood, context, etc, see Awtrey, Thornley, Dannals, Barnes, & Uhlmann, 2021; Kane, 1986).

One of the challenges to this rater context-driven perspective is that there is no clear guidance about the proportion of variability in rater effects that might contribute to reliability. This is because systematically varying work contexts are not typically included as part of the measurement design for performance ratings (e.g., Putka & Hoffman, 2014; Schmidt et al., 2000). If some portion of rater-related variance contributes to universe score, then the classification of all rater-related variance as a contribution to unreliability (e.g., Schmidt et al., 2000) will result in erroneously inflated estimates of rater variance. However, it is known that even highly trained raters, evaluating performance in standardized environments and required to focus exclusively on ratee performance, commit known failures (Jackson et al., 2016; Putka & Hoffman, 2013). Thus, the notion that naturalistic employee performance ratings are error-free is untenable. A reasonable take therefore suggests that some, but not *all* rater variance might be associated with reliability.

A specific line of research has suggested smaller rater effects than those previously estimated. This research area has focused on performance ratings in particular occupations, such as in healthcare, applied psychology, ergonomics, and occupational safety (Burke et al., 2011, 2006). Burke, Landis, and

Burke (2014) report that the measurement designs used in these occupations typically involve two raters who evaluate the same ratee at the same time and in the same context. The authors report higher reliabilities for such designs with "provisional" interrater reliability estimates of around .80 (p. 534). However, this still leaves open the possibility that ratings from different raters in different contexts might, in part, reflect perspectives that vary meaningfully. If context-varied effects are substantial and yet are treated wholly as contributing to unreliable variance, then the reliability of ratings might be underestimated.

ACs present a measurement design that includes varied work-relevant contexts. Research on ACs has explored the issue of contextual perspectives in detail as it pertains to rater sensitivity to changes in situational characteristics in the form of exercise effects (i.e., variance relating to different AC exercise contexts, see Lance, 2012; Lance, Lambert, Gewin, Lievens, & Conway, 2004). Two recent studies modeled rater (or assessor), exercise, and a multitude of other measurement design effects. This allowed for a statistically partialled perspective on the extent to which raters differentiated between AC exercise contexts (Jackson et al., 2016; Putka & Hoffman, 2013). The most conservative estimates from these studies suggested that whilst partialling idiosyncratic rater and other effects, between 33.51% and 38.10% of variance in AC ratings was attributable to the capacity for raters to identify differences between exercises. These estimates, based on assessor ratings evaluated within each exercise, provide initial insights into the expected proportion of rater variance that might indicate sensitivity to ratee performance in different work-relevant contexts. The Jackson et al. and Putka and Hoffman estimates partialled rater-related effects and thus aspects of possible rater bias.

Summary and knowledge gaps related to supervisor performance ratings

Theoretical development on performance ratings has focused on measurement structure (Greguras & Robie, 1998; Hoffman, Lance, Bynum, & Gentry, 2010; Lance, Teachout, & Donnelly, 1992; O'Neill et al., 2015). The prevailing perspective on supervisory performance ratings appears to be that their interrater reliability is low at around .52 and that this outcome is due to unreliability based on large rater-related effects (Schmidt & Hunter, 1996; Schmidt et al., 2000). However, a statistically partialled perspective on the measurement structure of supervisory performance ratings is currently unavailable. Such a perspective is required to add clarity to the literature on this widely applied measure.

To develop a theoretical understanding of the structure of supervisor ratings, a study is required that partials sources of variance central to their measurement design (raters, assessees, items, performance dimensions, and their interactions) whilst acknowledging the effects of aggregation. This leads to our first Research Question:

Research Question 1: On aggregation, what proportion of the variance in supervisory performance ratings is uniquely associated with: raters, assessees, items, performance dimensions, and their interactions?

Given previous research findings on the interrater reliability of performance ratings (LeBreton et al., 2014; Schmidt et al., 2000; Viswesvaran et al., 1996), we expect to find sizable rater effects in our results. Murphy and DeShon (2000a, 2000b) argue that at least some proportion of rater variance might contribute to universe score because raters evaluate ratees in different work contexts. Yet, it is highly unlikely that *all* rater variance contributes to universe score.

Although the measurement design of performance ratings does not typically differentiate between contextual influences, in contrast, ACs do differentiate between work contexts. Modeled in both the Jackson et al. (2016) and Putka and Hoffman (2013) estimates[4] was the potential for assessors to be sensitive to (a) performance within exercises and (b) performance on dimensions that vary by exercise

[4]In both studies, the estimate of between 33.51% and 38.10% of variance explained in AC ratings is based on the sum of the interaction between participants and exercises plus the interaction between participants, exercises, and dimensions.

(see also Hoffman, Kennedy, LoPilato, Monahan, & Lance, 2015). These findings could help to inform on the proportion of rater variance in supervisor ratings that is associated with sensitivity to ratee performance in different work contexts. The intention here is not to provide the definitive and final response to the question about which proportion of rater variance contributes to universe score. However, we seek to provide an approximation of the expected outcome when an informed proportion rater variance is accounted for by sensitivity to different performance contexts. This leads to our second Research Question:

Research Question 2: How do reliability estimates for performance ratings change when accounting for rater sensitivity across different performance contexts?

Results relating to our research questions will facilitate a consideration of how the multiple effects relevant to the measurement design of performance ratings contribute to universe score or unreliability. This consideration will, in turn, inform on the multifactor and components of the mediated theoretical models summarized by Murphy (2008).

Method

We reanalyzed subgroups from the data sets that appeared in Jackson et al. (2020). In the original study, the authors focused on a multisource measurement design. The emphasis of the current study is on supervisory ratings, for which there were two separate samples available in the Jackson et al. database. As a supplementary analysis and to test whether our findings replicated in different roles, we repeated our analyses on the other individual sources available in the data set. The Jackson et al. database allowed a unique level of complexity as it modeled the main features of the supervisory ratings measurement design (including items, dimensions, and raters) for data that potentially present a challenge for applied researchers to obtain. Data from two different samples were available for analysis. Each of these samples reflected a specific, albeit similar measurement design. However, each design was sufficiently different to offer insights about the potential for cross-sample generalization.

Sample 1

Participants

Participants in Sample 1 included 392 unique managerial ratees (298 men, 94 women) and 244 unique supervisory raters (183 men, 61 women) who were managers ranked a level above and who directly supervised ratees. Although supervisor ratings were our primary focus, we aimed to provide the reader with comparative findings from other roles available in the data set. We therefore included separate ratings from 420 direct reports (315 men, 105 women), 775 colleagues (581 men, 194 women), and 579 stakeholders (434 men, 145 women). The participant organization was involved in manufacturing in the United Kingdom. The main purpose for the procedure used in Sample 1 was for employee development. Neither ethnicity nor age data were collected out of concerns related to confidentiality.

Measurement design

All participant ratees (p) were assessed by raters (r, an average of 2 per role or source) who assessed on rating items (i, on average[5] 16.46 for each dimension), which were nested in performance dimensions (d, totaling 4). This design is typical of the type reported in the literature on performance ratings (e.g., Greguras & Robie, 1998; O'Neill et al., 2015).

[5]We applied harmonic mean values for averaging facet levels, in keeping with Brennan (2001)

Sample 2

Participants

Participants in Sample 2 included 342 unique managerial ratees (216 men, 126 women). The mean age of ratees in the full data set was 38.31 (SD = 9.65).[6] Ratees were assessed by 397 unique raters ranked a level above and who supervised ratees. As with Sample 1, direct supervisor ratings were our focus. However, we included for analysis data from other roles for comparison, including those from direct reports ($N = 833$), peers ($N = 872$), and clients ($N = 272$). Demographics on gender were only available for the total number of raters in the data set, including self-ratings (1579 men, 1057 women with a mean age of 40.24, SD = 9.89; note that 262 of these cases involved self-ratings, which we did not analyze). No further demographics were available. Unlike in Sample 1, the Sample 2 data set reflected ratings from different client organizations who made on-demand use of the performance management system. These client organizations were involved in banking, retail, accounting, insurance, human resources, and management consulting businesses in the United Kingdom. Applications of the procedure in Sample 2 depended on client requirements but included performance assessment and employee development.

Measurement design

In Sample 2, participant ratees (p) were assessed by raters (r, 2 on average per role or source) on rating items (i, on average 10.04 per performance dimension), which were nested in dimensions (d, total = 24). A mean of 5.40 dimensions were, in turn, nested in each of 5 summary dimension categories (c). The nested component of this measurement design related to dimensions was not present in Sample 1. Different clients made use of the performance management facility in Sample 2 to meet their specific demands, and so variation was apparent in the numbers of levels relating to sources of variance in this measurement design.

Rating procedures in samples 1 and 2

The procedures in both Samples 1 and 2 were developed on the basis of job analyses relating to the positions being evaluated (e.g., Williams & Crafts, 1997). Example rating items from Samples 1 and 2 respectively were: *"Ensures the strategy, objectives, and activities of the team are focused on addressing customer needs"* and *"Gives ongoing and constructive performance-related feedback."* In Sample 1 a rating scale was used ranging from 1 (the rater has never observed this behavior) to 5 (the rater always observes this behavior). In Sample 2, a percentile (0–100) score was available, which took the original 1 (strongly disagree) to 6 (strongly agree) scale rating and referenced this against responses from a norm group. Full definitions for the performance dimensions assessed in both Samples 1 and 2 appear in Appendix A1 of Jackson et al. (2020). These were described in the original study as "job-critical knowledge, skills, abilities, and other characteristics identified in the job analysis for each sample" (p. 318). Dimension titles appear in the Appendix of the present article (Table A3).

Rater training in Sample 1 covered use of the online platform used by the organization to input ratings and the use of mock assessments together with a discussion centered on a comparison of ratings. The latter was based on a frame-of-reference training procedure (e.g., Bernardin & Buckley, 1982). Sample 2 training involved a half-day course that covered procedural content and a mock assessment akin to that described for Sample 1. Training outcomes were not assessed by the organization in Sample 1. For Sample 2, training performance was assessed and only those who passed a training evaluation could proceed to use the rating procedure. The organization in Sample 1 used the evaluation on different occasions, but Jackson et al. (2020) were only provided access to ratings relevant to one evaluation period. The Sample 2 procedure was a one-off assessment.

[6]The mean and SD for age were based on ratings from all roles in Sample 2, including self-ratings. No other demographic information was available.

Data analysis

Measurement design, effects, and generalization

Both measurement designs in Samples 1 and 2 required the estimation of 11 separate effects each, although some of the specific effects in each sample were different to one another. The number of effects in the present study differs from that in the original because of the absence of a source effect and source-related interactions, given the focus here on supervisors. Full descriptions of the effects estimated in this study are provided in the Appendix, Tables A1 and A2. Briefly, across samples, we were able to simultaneously estimate effects associated with general performance (participant ratee main effects, akin to a general effect in classical test theory, CTT, or latent variable theory, LVT), Participant × Dimension interactions (akin to dimension-related effects or an indication of discriminant validity in CTT and LVT), multiple rater-related effects (e.g., CTT analogues of rater leniency or severity), and item-related effects. In Samples 1 and 2, items were nested in dimensions. A key feature of Sample 2 was that dimensions were nested in summary 2nd-order dimension categories.

Bayesian inference

Bayesian inference offers practical and statistical advantages over traditional approaches to estimation in the random effects models often applied as a basis for G theory (as detailed in Jackson et al., 2016, 2020; LoPilato, Carter, & Wang, 2015). We therefore opted to apply Bayesian inference to our data and, in doing so, we respond to general calls in the literature to explore applications of Bayesian statistics in applied psychology (Kruschke, Aguinis, & Joo, 2012; Zyphur, 2009; Zyphur, Oswald, & Rupp, 2015).

Ill-Structured designs and aggregation

Raters and ratees were neither fully crossed, but nor were they perfectly nested in any of the samples in this study. Thus, there was some degree of overlap between raters and ratees in both samples, constituting what is often referred to as an *ill-structured measurement design* (Putka, 2011). To address the data sparseness associated with ill-structured data configurations, we fitted a hierarchical Bayesian model (Gelman & Hill, 2007). An advantage of applying this approach is that it does not require that any data are deleted to fulfil the aim of developing a crossed design for the purposes of analysis. To reflect the degree of overlap between raters and ratees, we rescaled rater-related variance estimates in both samples using the *q*-multiplier approach (see Putka, Le, McCloy, & Diaz, 2008 for details). Moreover, we tailored formulae from the G theory literature (Brennan, 2001; Putka & Hoffman, 2014) and rescaled variance estimates to reflect aggregation across (a) rating items to form dimension scores, (b) dimension scores aggregated across raters, and (c) all items, dimensions, and raters to form overall scores. These formulae were applied to the posterior distributions of the model parameters so that we could obtain posterior distributions for all estimates.

Model specification

We used R 3.6.0 (R Core Team, 2019), Stan 2.19.1 (Stan Development Team, 2019), and brms 2.13.5 (Bürkner, 2017, 2018) to conduct the analyses in this paper. Samples 1 and 2 were configured with 11 variance components and 1 fixed intercept (to address our Research Question 1). Cross-sample comparability was facilitated by scaling each raw dataset to 1 standard deviation. This approach and others we have applied here assume that our data distributions approximated normality. Recent research on performance ratings challenges this assumption (Aguinis, Ji, & Joo, 2018). However, perusal of density and QQ plots did not raise concerns about appreciable deviations from normality in any sample relevant to the present work.

We applied weakly informative priors in all our analyses. For the fixed intercept, this was specified as a normal distribution with a mean of 3.06 (for Sample 1) and 3.05 (Sample 2) and a scale of 5.00 standard deviations. These mean values were selected based on rounding the mean of the dependent variable. For the standard errors of the random effects and the residual, we used the brms default

weakly informative prior of a student *t*-distribution with 3 degrees of freedom, a 0 mean, and a scale of 2.5 (Bürkner, 2017, 2018). The reasoning behind using these priors is that they will not allow the analysis to return values that are conceptually impossible. Whilst, at the same time, they are flexible to the extent they can permit a large range of values, even if the probability of them occurring is small. Weakly informative priors constitute the recommended practice for G theory models and have been successfully applied in organizational contexts involving raters (Jackson et al., 2016, 2020).

Simulations were conducted with four chains and with 10,000 iterations per chain. We treated the first 5,000 iterations as warm-up and retained the remaining chains for the main analysis. Convergence was acceptable in all analyses, according to visual inspections of trace, density, and autocorrelation plots. These outputs suggested good mixing of chains and did not raise any concerns about autocorrelation. Other indicators of effective convergence such as the scale reduction factor, effective sample size, and Monte Carlo standard errors were found to fall within acceptable parameters (see Gelman & Rubin, 1992).

Generalizability coefficients and rater sensitivity across work situations

On rescaling with the *q*-multiplier (as described in Putka et al., 2008), we estimated generalizability coefficients (G coefficients, Shavelson & Webb, 1991) for three types of generalization. First, we estimated generalization across different raters (generalization to r). This approach considers rater-related variance to be nonsystematic and to contribute to unreliability. Second, we estimated generalization to both different raters and rating items (generalization to i,r). This considers rater- and item-related variance to contribute to unreliability. Both generalization to r and i,r are consistent with the dominant perspective in the discipline, which considers rater-related variance to be classed as a contribution to unreliable variance (e.g., Schmidt & Hunter, 1996; Schmidt et al., 2000; Viswesvaran et al., 1996).

Third, we estimated reliability in keeping with the possibility raised by Murphy and DeShon (2000a) that at least some portion of rater-related variance might represent meaningful, systematic variation (see Research Question 2). We approximated the proportion of this potentially meaningful rater variance by referring to the AC literature, as detailed previously. Following the course of action suggested in Putka and Hoffman (2014), we reapportioned only systematic rater-related variance in our study. For the Jackson et al. (2016) AC estimate, we took the sum total of all systematic rater-related variance in the present study and partitioned it into 33.52% universe score and 66.48% unreliable variance. We repeated this principle for the Putka and Hoffman (2013, 38.10% universe score and 61.90% unreliable variance) estimate. We then used this approach as a basis for projected G coefficients for generalization to r only. As an aside and to provide clarity, in all G coefficient estimates we present, undifferentiated residual variance, which includes residual rater-related variance, was always specified, in full, as contributing to unreliability. We did not reapportion residual variance.

Results

Sample 1: supervisor ratings

Table 1 shows all 11 effects estimated for the supervisory ratings in Sample 1. Of these effects, 9 were relevant to comparisons between assessees (i.e., between-participant comparisons) and so constitute our focus. This is because in performance management, interest generally lies in how performance compares across different ratees. The results in Table 1 are presented initially in their pre-aggregated form. This is followed by estimates for aggregation across items to arrive at dimension scores, dimensions aggregated across raters, and overall scores across all raters, items, and dimensions. The aggregated presentation of results is likely to be relevant to many or most applications of performance ratings.

Table 1. Generalizability study at different levels of aggregation for supervisor ratings: sample 1.

	Pre-aggregation			Dimensions			Dimensions across raters			Overall ratings		
Effects	VC	Total Var %	BP Var %	Formula	VC	BP Var %	Formula	VC	BP Var %	Formula	VC	BP Var %
BP												
p	.1192	12.22	13.64	p	.1192	28.76	p	.1192	44.06	p	.1192	49.72
pd	.0030	.31	.34	pd	.0030	.73	pd	.0030	1.11	pd/n_d	.0008	.31
pi:d	.0560	5.74	6.41	$pi:d/n_{i:d}$	.0034	.82	$pi:d/n_{i:d}$	.0034	1.26	$pi:d/n_in_d$	.0002	.08
pr	.1414	14.49	16.18	pr	.1414	34.10	pr/n_r	.0701	25.89	pr/n_r	.0701	29.22
prd	.0285	2.92	3.26	prd	.0285	6.87	prd/n_r	.0141	5.21	prd/n_rn_d	.0035	1.47
pri:d	.3749	38.43	42.91	$pri:d/n_{i:d}$	.0228	5.50	$pri:d/n_rn_{i:d}$	.0113	4.17	$pri:d/n_rn_in_d$	.0007	.28
r	.0908	9.31	10.39	r	.0908	21.91	r/n_r	.0450	16.63	r/n_r	.0450	18.77
rd	.0019	.20	.22	rd	.0019	.47	rd/n_r	.0010	.35	rd/n_rn_d	.0002	.10
ri:d	.0580	5.95	6.64	$ri:d/n_{i:d}$	.0035	.85	$ri:d/n_{i:d}$	.0035	1.30	$ri:d/n_rn_in_d$	.0001	.04
Non-BP												
d	.0200	2.05										
i:d	.0818	8.38										
Generalization across:												
r			.20			.30			.46			.50
i,r			.14			.29			.45			.50
Projected generalization across r:												
Estimate 1			.34			.55			.65			.69
Estimate 2			.33			.52			.63			.67

Descriptions of the effects listed above are provided in the Appendix. p = participant ratee, d = performance dimension (or competency), i = rating item, r = rater. BP = between-participant, VC = variance component, Var = variance. G to = generalization across the effects that follow (for example, G to r = the expected generalizability coefficient when generalizing across different raters). All rater effects were corrected with the *q*-multiplier described in Putka et al. (2008). Estimate 1 based on Putka and Hoffman (2013, p. 38.10% contextual variance). Estimate 2 based on Jackson et al. (2016, p. 33.52% contextual variance). Observed G coefficients generalizing to r are given by the ratio of p + pd + pi:d to total BP variance. Observed G coefficients for generalizing to i,r are given by the ratio of p + pd to total BP variance. Projected Estimate 1 is given by the ratio of p + pd + pi:d + (38.10% of pr + prd + r + rd + ri:d) to p + pd + pi:d + pr + prd + r + rd + ri:d + pri:d. Projected Estimate 2 is given by the ratio of p + pd + pi:d + (33.52% of pr + prd + r + rd + ri:d) to p + pd + pi:d + pr + prd + r + rd + ri:d + pri:d. Note that pri:d is the estimate for residual variance and therefore does not contribute to universe score in projected estimates.

With reference to our Research Question 1, Table 1 shows a consistent pattern of results across the three different aggregation types relevant to this analysis. The assessee or participant main effect, σ_p^2, akin to a general effect, explained a large portion of variance across the dimension, dimension across raters, and overall aggregation types (28.76%, 44.06%, and 49.72%, respectively). Prominent across aggregation types were effects relating to raters. The Participant × Rater, σ_{pr}^2, interaction explained between 25.89% and 34.10% of variance. Likewise, the main effect for raters, σ_r^2, explained a substantial proportion of variance (between 16.63% and 21.91%). Collectively, rater-related effects explained most of the variance when aggregating to dimensions or to dimensions across raters (69.69% and 53.57% respectively) and around half at the overall aggregation level (49.88%).

We found that performance dimensions in Sample 1 explained a very small proportion of variance. The maximum contribution offered by the Participant × Dimension (σ_{pd}^2) effect was at the dimension-across-rater level of aggregation at 1.11% of the variance in ratings. The assessee Participant × Item nested in Dimension interaction ($\sigma_{pi:d}^2$) explained similarly low proportions of variance (≤1.26%).

This leads to a consideration of the G coefficients for Sample 1, which are presented in Table 1 for two types of generalization: specifically, to different raters (r), or items and raters (i,r). When generalizing to r or i,r, results were almost identical, given the large rater- and relatively small item-related effects evident in Table 1. G coefficients were uniformly low when attempting to generalize across different raters or items and raters (between .29 and .50).

Table 2. Generalizability studies aggregated to dimensions across raters for Non-Supervisor ratings: sample 1.

	Direct reports		Colleagues		Stakeholders	
Effects and Formula	VC	BP Var %	VC	BP Var %	VC	BP Var %
BP						
p	.1962	49.34	.1069	34.63	.1304	40.88
pd	.0022	.56	.0024	.78	.0046	1.45
pi:d/$n_{i:d}$	.0015	.39	.0023	.73	.0021	.67
pr/n_r	.1058	26.60	.0919	29.76	.0806	25.29
prd/n_r	.0116	2.93	.0191	6.19	.0160	5.03
pri:d/$n_r n_{i:d}$	.0088	2.21	.0103	3.35	.0094	2.95
r/n_r	.0627	15.77	.0720	23.33	.0704	22.09
rd/n_r	.0063	1.59	.0017	.56	.0026	.82
ri:d/$n_{i:d}$	.0024	.61	.0020	.65	.0027	.84
G to						
r		.50		.36		.43
i,r		.50		.35		.42

Descriptions of the effects listed above are provided in the Appendix. p = participant ratee, d = performance dimension (or competency), i = rating item, r = rater. BP = between-participant, VC = variance component, Var = variance. G to = generalization across the effects that follow (for example, G to r = the expected generalizability coefficient when generalizing across different raters). All rater effects were corrected with the *q*-multiplier described in Putka et al. (2008).

Sample 1: Non-Supervisor ratings

For brevity, we only reported a single aggregation level for non-supervisory ratings, namely that across dimensions and raters. Table 2 shows three different non-supervisory roles for Sample 1, including direct reports, colleagues, and stakeholders. The profile of variance was similar, regardless of role type, and was a similar type of profile to that observed for supervisors in Table 1. Across all three roles, σ^2_p (between 34.63% and 49.34%), σ^2_{pr} (between 25.29% and 29.76%), and σ^2_r (between 15.77% and 23.33%) all suggested prominent effects. Small effects were observed relating to dimensions, including σ^2_{pd} (≤1.45%) and $\sigma^2_{pi:d}$ (≤ .73%). G coefficients for non-supervisory roles were similar to those described above for the supervisory role. When generalizing to r or i,r, G coefficients were low (≤ .50).

Sample 2: supervisor ratings

Results for Sample 2 supervisor ratings are presented in Table 3. Despite reflecting a somewhat different measurement design, and in reference to our Research Question 1, the results in Sample 2 were similar to those observed in Sample 1. Table 3 shows that on aggregation, the primary contributions to variance in ratings were associated with a participant main effect σ^2_p (between 27.57% and 46.34%) and a rater-related effect ($\sigma^2_{r:p}$, between 38.50% and 50.68%). The second-order dimension effect (σ^2_{pc}) at the dimension and dimension-across-rater levels of aggregation explained 6.38% and 8.15%, respectively, of variance in ratings. Also, at the same levels of aggregation, the first-order dimension effect ($\sigma^2_{pd:c}$) was estimated at 2.11% and 2.70% and the item-nested-in-dimension effect ($\sigma^2_{pi:d:c}$) at 2.22% and 2.84%. However, gains in dimension-related variance in Sample 2 did not result in improved G coefficients when generalizing to raters. This is because, relative to dimension effects, rater effects were much larger. As found in Sample 1, G coefficients in Sample 2 were low when generalizing to r and i,r (≤ .49), regardless of aggregation type.

Sample 2: Non-Supervisor ratings

Table 4 shows outcomes for the three roles relevant to Sample 2, including direct reports, peers, and clients. As with the equivalent analysis in Sample 1, we only reported results for scores aggregated across items to form dimension scores and across raters. Findings for non-supervisory ratings were consistent with those for the supervisory ratings for Sample 2. The main contributors to variance in

Table 3. Generalizability study at different levels of aggregation for supervisor ratings: sample 2.

	Pre-aggregation			Dimensions			Dimensions across raters			Overall ratings		
Effects	VC	Total Var %	BP Var %	Formula	VC	BP Var %	Formula	VC	BP Var %	Formula	VC	BP Var %
BP												
p	.0721	8.26	9.46	p	.0721	27.57	p	.0721	35.20	p	.0721	46.34
pc	.0167	1.91	2.19	pc	.0167	6.38	pc	.0167	8.15	pc/n_c	.0033	2.15
pd:c	.0055	.63	.73	pd:c	.0055	2.11	pd:c	.0055	2.70	$pd:c/n_dn_c$	.0000	.03
pi:d:c	.0957	10.96	12.57	$pi:d:c/n_{i:d}$	.0058	2.22	$pi:d:c/n_{i:d}$	.0058	2.84	$pi:d:c/n_in_dn_c$	.0000	.01
r:p	.1215	13.92	15.96	r:p	.1215	46.48	$r:p/n_r$	.0788	38.50	$r:p/n_r$	.0788	50.68
r:pc	.0093	1.06	1.22	r:pc	.0093	3.55	$r:pc/n_r$	.0060	2.94	$r:pc/n_rn_c$	.0012	.77
r:pd:c	.0040	.46	.53	r:pd:c	.0040	1.53	$r:pd:c/n_r$	.0026	1.27	$r:pd:c/n_rn_dn_c$	.0000	.01
r:pi:d:c	.4367	50.04	57.35	$r:pi:d:c/n_{i:d}$	.0265	10.15	$r:pi:d:c/n_{i:d}n_r$	.0172	8.41	$r:pi:d:c/n_rn_in_dn_c$	.0000	.02
Non-BP												
c	.0099	1.14										
d:c	.0067	.77										
i:d:c	.0946	10.84										
Generalization across:												
r			.12			.36			.46			.49
i,r			.12			.34			.43			.48
Projected generalization across r:												
Estimate 1			.32			.58			.65			.66

Descriptions of the effects listed above are provided in the Appendix. p = participant ratee, d = performance dimension (or competency), i = rating item, r = rater, c = summary dimension category. BP = between-participant, VC = variance component, Var = variance. G to = generalization across the effects that follow (for example, G to r = the expected generalizability coefficient when generalizing across different raters). All rater effects were corrected with the *q*-multiplier described in Putka et al. (2008). Estimate 1 based on Putka and Hoffman (2013, p. 38.10% contextual variance). Estimate 2 based on Jackson et al. (2016, p. 33.52% contextual variance). Observed G coefficients generalizing to r are given by the ratio of p + pc + pd:c + pi:d:c to total BP variance. Observed G coefficients for generalizing to i,r are given by the ratio of p + pc + pd:c to total BP variance. Projected Estimate 1 is given by the ratio of p + pc + pd:c + pi:d:c + (38.10% of r:p + r:pc + r:pd:c) to p + pc + pd:c + pi:d:c + r:p + r:pc + r:pd:c + r:pi:d:c. Projected Estimate 2 is given by the ratio of p + pc + pd:c + pi:d:c + (33.52% of r:p + r:pc + r:pd:c) to p + pc + pd:c + pi:d:c + r:p + r:pc + r:pd:c + r:pi:d:c. Note that r:pi:d:c is the estimate for residual variance and therefore does not contribute to universe score in projected estimates.

Table 4. Generalizability studies aggregated to dimensions across raters for Non-Supervisor ratings: sample 2.

	Direct reports		Peers		Clients	
Effects and Formula	VC	BP Var %	VC	BP Var %	VC	BP Var %
BP						
p	.0585	30.98	.0622	33.88	.0751	38.79
pc	.0090	4.78	.0116	6.29	.0057	2.92
pd:c	.0039	2.09	.0027	1.47	.0015	.76
$pi:d:c/n_{i:d}$	.0050	2.66	.0051	2.75	.0038	1.94
$r:p/n_r$	.0853	45.18	.0733	39.90	.0806	41.63
$r:pc/n_r$	.0079	4.17	.0090	4.89	.0050	2.56
$r:pd:c/n_r$	.0014	0.77	.0015	.84	.0037	1.90
$r:pi:d:c/n_{i:d}n_r$	.0177	9.39	.0183	9.97	.0184	9.48
G to						
r		.38		.42		.42
i,r		.36		.40		.42

Descriptions of the effects listed above are provided in the Appendix. p = participant ratee, d = performance dimension (or competency), i = rating item, r = rater, c = summary dimension category. BP = between-participant, VC = variance component, Var = variance. G to = generalization across the effects that follow (for example, G to r = the expected generalizability coefficient when generalizing across different raters). All rater effects were corrected with the *q*-multiplier described in Putka et al. (2008).

ratings across all rater roles were σ^2_p (between 30.98% and 38.79%) and $\sigma^2_{r:p}$ (between 39.90% and 45.18%). The contribution of dimension-related variance was small but differed somewhat across roles and was primarily associated with the second-order dimension effect σ^2_{pc} (ranging from 2.92% with clients, up to 6.29% with peers). G coefficients were once again low when generalizing to different raters and different items as well as raters (≤ .42).

Projections based on reapportioned systematic rater variance

Entries for projected generalization across r, based on a reapportioning of systematic rater variance guided by the AC literature, appear in Tables 1 and 3 (see Research Question 2) for supervisor ratings. The results of this approximation were similar, regardless as to whether the Putka and Hoffman (2013, Estimate 1) or Jackson et al. (2016, Estimate 2) estimates were applied. Reliability increased substantially when systematic rater-related variance was reallocated according to the AC-based estimates. At the overall level of aggregation in Sample 1, reliability increased from the original estimate of .50 to a maximum of .69. At the overall level of aggregation in Sample 2, reliability increased from .49 to a maximum of .66. These projected estimates still do not meet criteria ordinarily set for acceptable reliability (Lance et al., 2006; LeBreton et al., 2014). However, they do move the reliability estimates closer to these criteria.

Discussion

The theoretical development of performance ratings has focused on their measurement structure relating particularly to general performance (Scullen, Mount, & Goff, 2000), performance dimensions (Borman & Brush, 1993; Kenny & Berman, 1980), and rater effects (Lance et al., 1992). Murphy (2008) described three models to explain the relationship between the performance and performance ratings, including *one-factor, multifactor*, and *mediated* models. Many current estimates of the measurement structure of performance ratings refer to the one-factor model based on classical test theory, where rater-related variance is typically assumed to contribute to unreliability (e.g., Viswesvaran et al., 1996). Less attention has been directed towards the multifactor and mediated models and the related possibility that at least some systematic rater-related variance might contribute to universe score. A statistically partialled evaluation of the measurement design usually described for performance ratings would inform these perspectives. It is this partialled account of the measurement structure of performance ratings that we sought to present (see Research Question 1). We further considered the impact of different perspectives on what defines multiple sources of universe score and unreliability in ratings, particularly regarding the status of systematic rater-related effects (Murphy & DeShon, 2000a, 2000b, see Research Question 2).

Our findings suggest that the structure of supervisor ratings tends to primarily reflect general performance and rater effects. This structure held across three different types of aggregation and two different measurement design variations. The largest portion of variance associated with raters in both samples was an effect involving both raters and participant ratees (σ^2_{pr} in sample 1 and $\sigma^2_{r:p}$ in sample 2). This implies that different raters held varying perspectives on ratee performance. When rater effects were treated as contributing to unreliable variance, as in the one-factor model, and in keeping with results from previous studies (LeBreton et al., 2014; Rothstein et al., 1990; Schmidt et al., 2000; Viswesvaran et al., 1996), we found low reliability estimates for supervisor ratings (≤ .50 for overall aggregation). The only contribution of note to universe score was that associated with a general performance effect. Our results further indicate that the reliability of performance ratings is undermined by the relatively small contribution of dimension effects (<3% of variance explained for overall aggregation).

Murphy and Deshon (2000a, 2000b) suggest that some proportion of rater variance might present meaningfully different context-based perspectives on a ratee. In our study, we estimated this proportion based on findings from the AC literature (Jackson et al., 2016; Putka & Hoffman, 2013). Our

results suggested that even when using conservative estimates, projected reliabilities increased substantially (from .50 to a maximum of .69 in Sample 1 and from .49 to .66 in Sample 2 for overall ratings) when systematic rater-related variance was reallocated according to AC-based estimates (see Research Question 2). These increases did not result in outcomes that met acceptability criteria often applied to reliability coefficients (Lance et al., 2006; LeBreton et al., 2014). Nonetheless, they approached such criteria (at between .66 and .69), and our estimates were based on the most conservative figures available in the Jackson et al. and Putka and Hoffman studies.

A statistically partialled perspective on the structure of performance ratings

By simultaneously partialling for all systematic effects relevant to their measurement design (see Tables 1 and 3), we suggest new insights into the structure of supervisory ratings. We were unable to find more detailed treatments of reliability in performance ratings in the literature, with previous studies being based on separate intra and interrater reliability estimates (e.g., Viswesvaran et al., 1996), or on simultaneously modeled but incomplete effects (e.g., Greguras & Robie, 1998; O'Neill et al., 2015).

In response to our Research Question 1 across 2 samples, it was clear on aggregation that the structure of supervisory performance ratings was primarily concerned with (a) person main effects (also referred to as general performance, σ^2_p, >27% of the variance in ratings) and (b) rater-related effects (various main effects and interactions involving raters and ratees, >49%, see Tables 1 and 3). The main contributions to variance in supervisor ratings were therefore associated with a general, positive-manifold-type appraisal of ratee performance (Ree, Carretta, & Teachout, 2015; Viswesvaran, Schmidt, & Ones, 2005), coupled with interactions involving participant ratees and raters (Lance et al., 1992). Similar effects were apparent in the other organizational roles we tested for comparison (see Tables 2 and 4).

Across both samples, our findings suggest that performance dimensions contributed only small proportions of variance ($\sigma^2_{pd} \leq 2.70\%$) to the structure of performance ratings. Our estimate of the contribution of these performance dimensions was somewhat lower than the $\sigma^2_{pd} \approx 6\%$ of variance estimated in O'Neill et al. (2015), where item-related effects were not modeled. In our Sample 2, we modeled the analogue of 2nd-order dimensions, which, as found in other contexts (see Hoffman, Melchers, Blair, Kleinmann, & Ladd, 2011), explained greater proportions of variance than our analogue of 1st-order dimensions (i.e., σ^2_{pd}). However, even these effects were too small ($\sigma^2_{pc} \leq 8.15\%$) to have any appreciable influence on reliability outcomes.

The question of rater-related variance

Uncertainty is apparent in the literature with respect to the status of rater-related variance. The relevant body of literature is silent on precisely what proportion of rater-related variance should be specified as contributing to universe score in supervisor ratings. Some proportion of rater-related variance might indeed represent meaningful context-perspective effects, as has been suggested from a conceptual stance (Murphy & DeShon, 2000a, 2000b; Putka et al., 2014). But surely not all rater variance could be reasonably classified as universe score. For example, differences based on personality, mood, role, cognitive ability, or any number of other characteristics could be relevant to rater-related variance. In the absence of other specific guidance about the reasons for rater-related variability, a conservative course of action is to treat all rater-related variance as a contribution to unreliability. It is this conservative approach that prevails in the research literature (see LeBreton et al., 2014 for a summary).

The projected estimates for reliability we present based on knowledge from the AC literature provide a step toward finding some, informed mid-way between two hypothetical extremes (i.e., all rater variance = unreliable variance, versus all rater variance = universe score). We provide what can

be thought of as a theoretical estimate and not one that is intended to deliver the definitive answer to the question about what proportion of rater-related variance is universe score. Locating that precise figure might present a difficulty for the discipline, given that the definition of rater context perspectives might differ markedly by sample. Moreover, the measurement design of performance ratings represented in the literature does not typically (or possibly ever) include an estimate of rater-related context. This is likely because such contexts are likely to change unsystematically in real-world scenarios.

Comparison with the multisource rating design

Our study offers a unique opportunity to compare our results, focused on supervisor ratings, directly with multisource ratings, given that we reanalyzed data from a multisource ratings data set. The conclusion in the original study was that multisource ratings showed encouraging evidence for reliability, even when all systematic rater-related variance was treated as contributing to unreliability (≥ .81). But using the same data sets and, like in the original study, treating systematic rater variance as contributing to unreliability, we found considerably lower reliabilities for supervisor ratings (≤ .50).

The reason for this apparent discrepancy is due to the presence of a source effect in the multisource design that is absent in the supervisory ratings design.[7] Relative to source effects in Jackson et al., rater effects were small. However, when the source effects were removed, as in the present study, rater effects became more prominent relative to other remaining effects in the supervisory ratings design. This finding highlights the relative nature of the effects that contribute to reliability estimation. The addition of even one substantial effect in a measurement design can make a sizable difference to a reliability estimate.

Implications for researchers and practitioners

Murphy (2008), in his description of multifactor and mediated models, suggested that performance is only one of several possible components that contribute to performance ratings. Our findings suggest that universe score components of performance ratings are defined primarily by general effects. General effects could partly represent general performance but could also represent individual differences on psychological constructs (Putka & Hoffman, 2013, 2014; Ree et al., 2015). We found only a small portion of variance in the component of the measurement design that clearly attempts to formalize an evaluation of performance in the form of performance dimensions. In comparison to multisource ratings, in supervisor ratings there is a greater reliance on a smaller number of effects typically deemed to contribute to reliability (specifically, general effects and performance dimensions). Putting aside a consideration of rater-related effects, our results suggest that the reliability of supervisor ratings is primarily reliant on a relatively large general effect, particularly given that dimension effects tend to be small.

The finding that dimensions (or competencies) contributed small proportions of variance in our study is consistent with findings in other settings (e.g., ACs, interviews, and situational judgment tests, see Jackson et al., 2016, 2020; Lance et al., 2004; O'Neill et al., 2015; Putka & Hoffman, 2013). We believe that researchers could address this phenomenon in future studies. Small dimension effects might be a consequence of conceptual issues (e.g., dimensions are not defined in ways that suit what it is that raters are able or prefer to evaluate) or due to time-related pressures and practicalities. For example, managers have a limited period at their disposal in which to complete appraisal forms and so they might only provide a similar rating across all dimensions. Yet another consideration is the number of occasions over which dimensions are evaluated, as we discuss below in our limitations section.

[7]We only generalize to r here for brevity and, in any case, the results for generalization to r and i,r were almost identical.

Table 5. Projected corrections for performance criterion unreliability by meta-Analytic study.

				Correction based on r_{yy}		
	K	*N*	r_{xy}	.52[a]	.66[b]	.69[c]
Cognitive ability						
Salgado, Anderson, Moscoso, Bertua, and de Fruyt (2003)	93	9,554	.29	.40	.36	.35
Bertua, Anderson, and Salgado (2005)	12	2,469	.22	.31	.27	.26
Employment interviews						
Huffcutt, Culbertson, and Weyhrauch (2014), pp. – structured	69	4,795	.35	.48	.43	.42
Huffcutt et al. (2014), pp. – unstructured	23	2,594	.12	.16	.15	.14
Assessment centers						
Hermelin, Lievens, and Robertson (2007)	27	5,850	.17	.24	.21	.20
Hardison and Sackett (2007)	49	4,198	.20	.28	.25	.24
Conscientiousness						
Salgado (2003), pp. – FFM	90	19,460	.17	.24	.21	.20
Salgado (2003), pp. – non-FFM	36	5,874	.11	.15	.14	.13
Biographical data						
Speer et al. (2021), pp. – empirically-keyed	49	20,564	.31	.44	.38	.37
Speer et al. (2021), pp. – rationally-keyed	22	16,279	.17	.24	.21	.20

Meta-analytic estimates above are based on Sackett, Zhang, Berry, and Lievens (2021). *K* = number of samples; *N* = number of participants across samples. r_{xy} mean meta-analytic predictor-criterion correlation. r_{yy} = .52 from Viswesvaran et al. (1996), r_{yy} = .66 from Sample 2 in the present study with aggregation to overall ratings and based on partitioning of rater-related variance from Putka and Hoffman (2013) and Jackson et al. (2016). r_{yy} = .69 from Sample 1 in the present study with aggregation to overall ratings and based on partitioning of rater-related variance from Putka and Hoffman (2013). r_{yy} = .66 to r_{yy} = .69 represents the range of r_{yy} estimates in this study post reallocation of rater-based variance.

We estimated the proportion of rater variance that might contribute to universe score by drawing on findings from the AC literature. In Table 5, we show corrections for meta-analytic r_{xy} validity coefficients corrected for the traditional .52 r_{yy} reliability estimate from Viswesvaran et al. (1996). We show the same type of corrections in Table 5 for the range of projected reliability estimates from the present study. Several of these corrections make a difference of note. For example, when corrected for r_{yy} = .52, r_{xy} for empirically-keyed biographical data = .44. When corrected for r_{yy} = .69, the same r_{xy} = .37.

Limitations

Partly because of the complexity of the models involved in this paper, we opted for an approach based on random effects models with Bayesian inference. This is not, however, the only approach that can be used to generate variance estimates, and confirmatory factor analytic (CFA) models can be used to address the same types of data structure and have the advantage of providing more detail about specific constructs of interest (Le, Schmidt, Harter, & Lauver, 2010). Despite these advantages, models with a large number of effects can be computationally impractical to analyze with CFA. Furthermore, CFA provides no straightforward approach toward handling ill-structured measurement designs (LoPilato et al., 2015; Putka et al., 2008). In contrast, random effects models based on Bayesian inference provide the capacity to handle a large number of effects and ill-structured measurement designs. Our random effects models moreover provided the level of detail required for us to address our research questions.

The samples in the present study were sizable and utilized measurement designs that were likely comparable. However, they also presented differences in their measurement designs for reasonable cross-sample comparisons. We found similar effects, not only across samples, but also across the supplementary roles. Essentially the same results were repeated across two samples, involving eight roles and two variations on a measurement design. That said, it would be helpful to investigate the unconfounded measurement design of performance ratings in a range of different types of occupation. For example, we do not know if our results will generalize to non-managerial samples.

Regarding our reapportioning of rater-based variance based on AC estimates, we were cognizant that, despite the similarities, there are differences in performance rating and AC procedures (e.g., AC exercises could represent maximal performance scenarios and ACs elicit performance in exercises that are likely designed to be different from one another). Accordingly, we applied the most conservative estimates from the most complex AC models we could find in the literature (i.e., from Jackson et al., 2016; Putka & Hoffman, 2013). We further reiterate that our intent is not to present the final word on the status of rater variance. However, we seek to provide an informed perspective on the possibility that even a moderate contribution of rater variance to universe score might make a difference to the estimated reliability of performance ratings. It is our hope that our estimates can be refined in future research.

It would be possible to test the Murphy and DeShon (2000a, 2000b) context-based perspective for rater variance directly with a quasi-experimental design. For example, raters (IV1, with >1 levels per ratee) could be assigned to systematically differing work contexts (IV2, >1 levels) but crossed such that all raters assess in all contexts. Rater effects could then be separated from contextual perspectives, with the former defined as unreliable variance and the latter defined as universe score. To address the potential for different raters to focus on different aspects of performance or to hold differing views on performance levels, standard-setting training could be introduced (e.g., Pulakos, 1986) as a third IV with 3 levels (trained, non-trained, and control). Scaled covariates could be introduced into the study (e.g., for rater personality, cognitive ability, mood, etc.). The design described here could present a potentially fruitful opportunity for future research and could help to provide further guidance on what portion of rater variance contributes to universe score. However, it might raise problems of generalization, as a reasonable take suggests aspects of many real-world work contexts routinely change unsystematically.

We found relatively small dimension effects in our study when compared to the general and rater-related effects in our models. It is possible that this finding was specific to the samples included in this study. However, the dimensions included in the original study (see Appendix Table A3) suggest at least some conceptual distinctions between the dimensions that were applied. The dimensions in our samples are reminiscent of those found in research guidance elsewhere (e.g., Arthur, Day, McNelly, & Edens, 2003). Moreover, our findings are consistent with those in other contexts (as mentioned previously).

On general performance, Murphy et al. (2019) note that shared variance among raters might not purely indicate ratee performance, but could also reflect nonperformance-related individual rater and system-related characteristics. These issues remain relevant, even if idiosyncratic rater effects are isolated, as they were in our study. The training procedures used in our samples were aimed at mitigating nonperformance effects. Nonetheless, the points that Murphy et al. raise remain as important background considerations when evaluating the meaning of evaluations generated by any measurement design employing external raters.

The data set in this study did not include repeated measures of the same group of ratees. Thus, we could not model the effect of occasions of measurement (Brennan, 2001 provides a discussion on this topic). It would be interesting to know, particularly when the aim is developmental in nature, how occasions interact with other effects in the performance ratings measurement design. For example, would the presence of an occasions facet increase the magnitude of effects associated with dimensions whilst considering the expectation that performance is expected to develop over time? One possible explanation for small dimension effects is that raters possibly require a greater number of opportunities to observe dimension-related behavior on different occasions. Nonetheless, as with multisource ratings, occasions are not typically described in the literature as being fundamental to the measurement design of performance ratings (e.g., Greguras & Robie, 1998; LeBreton et al., 2014; Murphy & DeShon, 2000a; O'Neill et al., 2015).

Conclusion

Our results suggest a measurement structure for supervisory ratings that primarily reflects general and rater-related effects, but with substantially smaller effects for performance dimensions. Our findings suggest that reallocating even a moderate portion of systematic rater-related variance to universe score

makes a sizable difference to reliability estimates for performance ratings. Future research could offer further insights into what proportion of rater variance is likely to be best classed as universe score. In addition, reliability gains in performance ratings would likely follow if it were possible to improve dimension-related evaluations.

Disclosure statement

For disclosure, the fifth and sixth authors had financial interests in the processes used in the present study. However, the remaining authors, who did not hold financial interests in these processes, were granted full liberty to analyze and interpret any associated data to help preserve impartiality. Only full data sets, not selected subsets of variables, were provided and analyzed.

References

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

Aguinis, H., Ji, Y. H., & Joo, H. (2018, Dec). Gender productivity gap among star performers in STEM and other scientific fields. *Journal of Applied Psychology*, *103*(12), 1283–1306. doi:10.1037/apl0000331

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology*, *56*(1), 125–154. doi:10.1111/j.1744-6570.2003.tb00146.x

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93*(2), 435–442. doi:10.1037/0021-9010.93.2.435

Awtrey, E., Thornley, N., Dannals, J. E., Barnes, C. M., & Uhlmann, E. L. (2021). Distribution neglect in performance evaluations. *Organizational Behavior and Human Decision Processes*, *165*, 213–227. doi:10.1016/j.obhdp.2021.04.007

Bartram, D. (2005). The Great Eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology*, *90*(6), 1185–1203. doi:10.1037/0021-9010.90.6.1185

Bennett, W., Jr., Lance, C. E., & Woehr, D. J. (2006). *Performance measurement: Current perspectives and future challenges*. Erlbaum. http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2006-07017-000&site=ehost-live

Bernardin, H. J., & Buckley, M. R. (1982). Strategies in rater training. *Academy of Management Review*, *6*, 205–212.

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology*, *78*(3), 387–409. doi:10.1348/096317905X26994

Borman, W. C., & Brush, D. H. (1993). More progress toward a taxonomy of managerial performance requirements. *Human Performance*, *6*(1), 1–21. doi:10.1207/s15327043hup0601_1

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

Burke, M. J., Landis, R. S., & Burke, M. I. (2014). 80 and beyond: Recommendations for disattenuating correlations. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *7*(4), 531–535. doi:10.1111/iops.12190

Burke, M. J., Salvador, R. O., Smith-Crowe, K., Chan-Serafin, S., Smith, A., & Sonesh, S. (2011). The dread factor: How hazards and safety training influence learning and performance. *Journal of Applied Psychology*, *96*(1), 46–70. doi:10.1037/a0021838

Burke, M. J., Sarpy, S. A., Smith-Crowe, K., Chan-Serafin, S., Salvador, R. O., & Islam, G. (2006). Relative effectiveness of worker safety and health training methods. *American Journal of Public Health*, *96*(2), 315–324. doi:10.2105/AJPH.2004.059840

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

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

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, *63*(1), 83–117. doi:10.1111/j.1744-6570.2009.01163.x

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.

DeNisi, A. S., & Murphy, K. R. (2017, Mar). Performance appraisal and performance management: 100 years of progress? *Journal of Applied Psychology*, *102*(3), 421–433. doi:10.1037/2Fapl0000085

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*(4), 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*(6), 960–968. doi:10.1037/0021-9010.83.6.960

Hardison, C. M., & Sackett, P. R. (2007). *Kriteriumsbezogene Validität des Assessment Centers: Lebendig und wohlauf?* [Criterion-related validity of assessment centers: Alive and well?]. H. Schuler Ed. Göttingen, Germany: Assessment Center zur Potenzialanalyse. Hogrefe.

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment*, *15*(4), 405–411. doi:10.1111/j.1468-2389.2007.00399.x

Hoffman, B. J., Kennedy, C. L., LoPilato, A. C., Monahan, E. L., & Lance, C. E. (2015, Jul). A review of the content, criterion-related, and construct-related validity of assessment center exercises. *Journal of Applied Psychology*, *100*(4), 1143–1168. doi:10.1037/a0038707

Hoffman, B. J., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology*, *63*(1), 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*(2), 351–395. doi:10.1111/j.1744-6570.2011.01213.x

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2014). Moving forward indirectly: Reanalyzing the validity of employment interviews with indirect range restriction methodology. *International Journal of Selection and Assessment*, *22*(3), 297–309. doi:10.1111/ijsa.12078

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

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

Kane, J. S. (1986). Performance distribution assessment. In Berk, R. A. (Ed). *Performance assessment: Methods & applications* (pp. 237–273). Baltimore, MD: Johns Hopkins University Press.

Kenny, D. A., & Berman, J. S. (1980). Statistical approaches to the correction of correlational bias. *Psychological Bulletin*, *88*(2), 288–295. doi:10.1037/0033-2909.88.2.288

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*(4), 722–752. doi:10.1177/1094428112457829

Kurz, R., & Bartram, D. 2002. Competency and individual performance: Modeling the world of work. In Robertson, I. T., Callinan, M., Bartram, D. (Eds.), *Organizational effectiveness: The role of psychology* (pp. 227–255). Chichester, UK: Wiley.

Lance, C. E. 2012. Research into task-based assessment centers. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 218–233). New York: Routledge/Taylor & Francis Group.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*(2), 202–220. doi:10.1177/1094428105284919

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*(2), 377–385. doi:10.1037/0021.9010.89.2.377

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*(4), 437–452. doi:10.1037/0021-9010.77.4.437

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin 87*(1) , 72–107.

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*(2), 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*(4), 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*(2), 692–717. doi:10.1177/0149206314554215

Murphy, K. R. (Ed.). (2003). *Validity generalization: A critical review*. Mahwah, NJ: Erlbaum. http://www.loc.gov/catdir/enhancements/fy0662/2002024474-d.html

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*(2), 148–160. doi:10.1111/j.1754-9434.2008.00030.x

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational and goal-based perspectives*. Thousand Oaks, CA: Sage.

Murphy, K. R., Cleveland, J. N., & Hanscom, M. E. (2019). *Performance appraisal and management*. Thousand Oaks, CA: Sage Publications.

Murphy, K. R., & DeShon, R. (2000a). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, *53*(4), 873–900. doi:10.1111/j.1744-6570.2000.tb02421.x

Murphy, K. R., & DeShon, R. (2000b). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology*, *53*(4), 913–924. doi:10.1111/j.1744-6570.2000.tb02423.x

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

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. *Organizational Behavior & Human Decision Processes*, *38*(1), 76–91. doi:10.1016/0749-5978(86)90027-0

Putka, D. J. (2011, April 16). Partitioning reliable and unreliable variance in dimension-exercise units.*Reevaluating assessment centers: New statistical approaches, new insights* 26th Annual Society for Industrial and Organizational Psychology Conference, Chicago, IL.

Putka, D. J., & Hoffman, B. J. (2013). Clarifying the contribution of assessee-, dimension-, exercise-, and assessor-related effects to reliable and unreliable variance in assessment center ratings. *Journal of Applied Psychology*, *98*(1), 114–133. 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., Hoffman, B. J., & Carter, N. T. (2014). Correcting the correction: When individual raters offer distinct but valid perspectives. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *7*(4), 543–548. doi:10.1111/iops.12193

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

R Core Team. (2019). *R: A language and environment for statistical computing*. In (Version 3.6.0). 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*(3), 409–427. doi:10.1017/iop.2015.16

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology*, *75*(2), 175–184. doi:10.1037/0021-9010.75.2.175

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2021). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*. doi:10.1037/apl0000994

Salgado, J. F. (2003). Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology*, *76*(3), 323–346. Journal of Occupational Psychology. doi:10.1348/096317903769647201.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & de Fruyt, F. (2003). International Validity Generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology*, *56*(3), 573–605. doi:10.1111/j.1744-6570.2003.tb00751.x

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

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology*, *53*(4), 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*(6), 956–970. doi:10.1037//0021-9010.85.6.956

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

Speer, A. B., Tenbrink, A. P., Wegmeyer, L. J., Sendra, C. C., Shihadeh, M., & Kaur, S. (2021). Meta-analysis of biodata in employment settings: Providing clarity to criterion and construct-related validity estimates. *Journal of Applied Psychology*. doi:10.1037/apl0000964

Stan Development Team. (2019). *Stan: A C++ library for probability and sampling*. In (Version 2.19.1). https://mc-stan.org

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25–29. doi:10.1037/h0071663

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*(5), 557–574. doi:10.1037/0021-9010.81.5.557

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*(1), 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.

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*(2), 387–389. doi:10.1177/0149206314549252

Appendix

Table A1. Guide to sources of variance in performance ratings for sample 1.

Effects	Brief description	Classic analogues[a]
p	Some participants (assessees) are generally rated higher than others, regardless of the dimension, rater, or item involved.	General performance, positive manifold
pd	Some participants are rated higher on some dimensions relative to others, regardless of the rater or item involved.	Dimension or trait effects
pi:d	Some participants are rated higher on some items (nested in dimensions) than others, regardless of the specific rater involved.	Intrarater effects, internal consistency, item-related effects
pr	Some participants are rated higher by some raters than by others, regardless of the items or dimensions involved.	Interrater halo effects
prd[a]	Some participants are rated higher on some dimensions than on others, regardless of the item involved. But this depends on who is rating them.	Interrater effects related to dimensions
pri:d	Highest-order effect confounded with and taken as an indication of residual variance.	Residual variance
r	Some raters provide higher ratings than others, but this could also affect participant rank-ordering because of the ill-structured measurement design.	General rater leniency, but affects participant rank ordering because of ill-structured design
rd	Some raters provide higher ratings that others on specific dimensions, regardless of item effects. But this could affect participant rank-ordering because of the ill-structured measurement design.	Rater leniency relating to specific dimensions, but affects participant rank ordering because of ill-structured design
ri:d	Some raters provide higher item-level ratings that others, regardless of specific dimensions. But this could affect participant rank-ordering given the ill-structured measurement design.	Rater leniency relating to specific items, but affects participant rank ordering because of ill-structured design
d	Some dimensions are rated higher on average than others	Average dimension scores
i:d	Some items (nested in dimensions) are rated higher than average than others	Average item ratings

p = participant assessee, d = performance dimension, i = rating item, r = rater. The last 2 effects in this table (i.e., d, i:d) do not interact with assessees and are therefore irrelevant to between-assessee comparisons. Rater-related effects are relevant to between-participant considerations in ill-structured measurement designs (Putka & Hoffman, 2013).

Table A2. Guide to sources of between-Participant variance in performance ratings for sample 2.

Effects	Brief description	Classic analogues[a]
p	Some participants (assessees) are generally rated higher than others, regardless of the dimension, rater, or item involved.	General performance, positive manifold
pc	Some participants are rated higher on some dimension categories relative to others, regardless of the rater or item involved.	Second-order trait effects
pd:c	Some participants are rated higher on some dimensions relative to others, regardless of the rater or item involved.	Dimension or trait effects
pi:d:c	Some participants are rated higher on some items (nested in dimensions, in turn, nested in categories) than others, regardless of the specific dimension or rater involved.	Intrarater effects, internal consistency, item-related effects
r:p	Ratees have specific groups of raters assigned. Some rater groups evaluate ratees higher than others, regardless of the items or dimensions involved.	Interrater halo effects
r:pc	Ratees have specific groups of raters assigned. Some rater groups evaluate ratees higher than others on categories, regardless of the items or dimensions involved.	Nested rater effects related to second-order dimensions
r:pd:c	Ratees have specific groups of raters assigned. Some rater groups evaluate ratees higher than others on dimensions (nested in categories), regardless of the items involved.	Nested rater effects related to dimensions
r:pi:d:c	Highest-order effect confounded with and taken as an indication of residual variance.	Residual variance
c	Some categories are rated higher on average than others	Average category scores
d:c	Some dimensions (nested in categories) are rated higher on average than others	Average dimension scores
i:d:c	Some items (nested in dimensions, nested in categories) are rated higher on average than others	Average item responses

p = participant assessee, d = performance dimension, i = rating item, r = rater, c = summary dimension category. The last 3 effects in this table (i.e., c, d:c, and i:d:c) do not interact with assessees and therefore are irrelevant to between-assessee comparisons. Rater-related effects are relevant to between-participant considerations in ill-structured measurement designs (Putka & Hoffman, 2013).

Table A3. Data transparency for archival data source: performance ratings.

Sample 1	Dimensions	p:s configuration (Jackson et al., 2020)	p configuration (Current study)
	Teamwork	X	X
	Organizational citizenship	X	X
	Results focused	X	X
	Motivation	X	X
Sample 2	Dimensions/ broad dimensions		
	Goal setting	X	X
	Delegating	X	X
	Independence	X	X
	Managing change	X	X
	Persuasive communication	X	X
	Project management	X	X
	Results orientation	X	X
	Organizational	X	X
	Attention to detail	X	X
	Commitment	X	X
	Information management	X	X
	Planning and organizing	X	X
	Interpersonal	X	X
	Communication skills	X	X
	Customer focus	X	X
	Developing others	X	X
	Interpersonal skills	X	X
	People management	X	X
	Team orientation	X	X
	Enterprise	X	X
	Leadership potential	X	X
	Motivation	X	X
	Resilience	X	X
	Risk-taking	X	X
	Self-confidence	X	X
	Strategy	X	X
	Analytic	X	X
	Creative	X	X
	Decision making	X	X
	Flexibility	X	X
	Problem solving	X	X
	Strategic awareness	X	X

X marks where data were used. In the original study, data were configured to a participant ratee (p) nested in source (s, i.e., p:s) configuration. This is contrasted against the present study, where we removed the source effect and configured our data with p as a crossed effect.