Journal of Applied Psychology © 2012 American Psychological Association

2013, Vol. 98, No. 1, 114 –133 0021-9010/13/$12.00 DOI: 10.1037/a0030887

Clarifying the Contribution of Assessee-, Dimension-, Exercise-, and Assessor-Related Effects to Reliable and Unreliable Variance in Assessment Center Ratings

Dan J. Putka

Human Resources Research Organization, Alexandria, Virginia

Brian J. Hoffman

University of Georgia

Though considerable research has evaluated the functioning of assessment center (AC) ratings, surprisingly little research has articulated and uniquely estimated the components of reliable and unreliable variance that underlie such ratings. The current study highlights limitations of existing research for estimating components of reliable and unreliable variance in AC ratings. It provides a comprehensive empirical decomposition of variance in AC ratings that: (a) explicitly accounts for assessee-, dimension-, exercise-, and assessor-related effects, (b) does so with 3 large sets of operational data from a multiyear AC program, and (c) avoids many analytic limitations and confounds that have plagued the AC literature to date. In doing so, results show that (a) the extant AC literature has masked the contribution of sizable, substantively meaningful sources of variance in AC ratings, (b) various forms of assessor bias largely appear trivial, and (c) there is far more systematic, nuanced variance present in AC ratings than previous research indicates. Furthermore, this study also illustrates how the composition of reliable and unreliable variance heavily depends on the level to which assessor ratings are aggregated (e.g., overall AC-level, dimension-level, exercise-level) and the generalizations one desires to make based on those ratings. The implications of this study for future AC research and practice are discussed.

*Keywords:* assessment center, assessors, reliability, variance components

*Supplemental materials:* [http://dx.doi.org/10.1037/a0030887.supp](http://dx.doi.org/10.1037/)

Over the past three decades, myriad studies have examined the latent structure of assessment center (AC) ratings. In gen- eral, this literature has produced two broad schools of thought with regard to AC functioning. One has emphasized the impor- tance of taking a construct/dimension-focused perspective to AC research and practice (e.g., Arthur, Day, & Woehr, 2008; Meriac, Hoffman, Woehr, & Fleisher, 2008; Rupp, Thornton, & Gibbons, 2008; Shore, Thornton, & Shore, 1990), and another has emphasized a task/exercise-focused view (e.g., Jackson,

2012; Jackson, Stillman, & Atkins, 2005; Lance, 2008; Lance et al., 2000; Lance, Lambert, Gewin, Lievens, & Conway, 2004; Lowry, 1997; Sackett & Dreher, 1982). In light of these per- spectives, as well as variability in AC design, implementation, and modeling strategies, much of this research has been devoted

This article was published Online First December 17, 2012.

Dan J. Putka, Human Resources Research Organization, Alexandria, VA; Brian J. Hoffman, Department of Psychology, University of Georgia. The first author presented an earlier version of this article at the 26th annual Society for Industrial and Organizational Psychology Conference, April 2011, Chicago, IL. We thank Kevin Murphy, Paul Sackett, Suzanne Tsacoumis, and Deborah Whetzel for their helpful comments on an earlier

version of this article.

Correspondence concerning this article should be addressed to Dan J. Putka, Human Resources Research Organization, 66 [Canal Center Plaza,](mailto:dputka@humrro.org) Suite 700, Alexandria, VA 22314-1591. E-mail: [dputka@humrro.org](mailto:dputka@humrro.org)

to methodological and design-based moderators of dimension and exercise effects (Arthur & Day, 2011; Lance et al., 2004; Lievens, 1998; Woehr & Arthur, 2003). As a whole, reasonable arguments are made from both perspectives, and the body of findings underscores the importance of moving beyond dimensions-only and exercises-only interpretations to more multifaceted, nuanced perspectives on AC functioning (e.g., Borman, 2012; Hoffman, 2012; Hoffman, Melchers, Blair, Kleinmann, & Ladd, 2011; Howard, 2008; Lievens, Chasteen, Day, & Christiansen, 2006).

Despite advances in the AC literature, a key element of AC functioning has been left woefully underexamined for a litera- ture that has matured to this point. Specifically, little research has been devoted to underlying components of reliable and unreliable variance in assessor ratings. This is despite the fact that reliability is a fundamental psychometric property upon which the quality of assessment scores should be judged and has implications for subsequent estimation and interpretation of validity evidence for AC scores (American Educational Re- search Association, American Psychological Association, & National Council on Measurement in Education, 1999; Society for Industrial and Organizational Psychology, 2003). Further- more, by estimating components that compose reliable and unreliable variance, one can gain a better understanding of how ACs function than what is revealed by simple reliability or validity coefficients (Cronbach, Gleser, Nanda, & Rajaratnam,

1972; Haertel, 2006; Putka & Sackett, 2010). As we discuss in

114

the current article, reliable and unreliable variance in assessor ratings is a function of myriad components, not simply the dimension and exercise effects that have tended to dominate the attention of AC researchers to date. Furthermore, the composi- tion of reliable and unreliable variance in assessor ratings is a function of the level to which such ratings are aggregated (e.g., overall AC score level, dimension-level, exercise-level), and the generalizations one desires to make regarding the resulting scores (Cronbach et al., 1972). Both of these issues have received little attention from AC researchers, yet are beneficial to examine because they have implications for prevailing sci- entific perspectives on AC functioning and practice.

**The Current Study**

In the current article, we attempt to fill several voids in the AC literature. First, we illustrate how common methodological ap- proaches to examining AC functioning have hampered the emer- gence of a clear conceptual and empirical treatment of the com- position of reliable and unreliable variance in assessor ratings by confounding distinct, substantively meaningful sources of vari- ance. Second, we define components of variance underlying as- sessor ratings and discuss limitations of the current literature for estimating their magnitude. With these conceptual foundations in place, we use random effects models to estimate components of reliable and unreliable variance using three samples of data from a high-stakes operational AC program used for promoting employ- ees to first-line supervisor positions. We then use the estimated components to illustrate how the composition of reliable and unreliable variance changes depending on the level to which assessor ratings are aggregated (e.g., overall AC score level, dimension-level, and exercise-level), and the generalizations one desires to make on the basis of those ratings. Finally, we highlight the implications of our findings for future AC research and prac- tice.

**A Surprisingly Silent Literature**

The AC literature’s silence on components of reliable and unreliable variance in assessor ratings can be traced to the ap- proaches used to model AC data. The dominant methodological approach in the AC literature has been grounded in Campbell and Fiske’s (1959) multitrait–multimethod (MTMM) framework and focused on issues of construct validity rather than reliability. Over the past three decades, MTMM-based studies have commonly involved fitting confirmatory factor analytic (CFA) models to assessor ratings of dimensions that are made at the conclusion of each exercise, or postexercise dimension ratings (PEDRs; Lance,

2008). Though typical CFA-based approaches have been criticized in light of convergence and admissibility issues, as well as the difficulty in using them to handle ill-structured ratings designs characteristic of ratings-based measures (e.g., Lance, Woehr, & Meade, 2007; Putka, Lance, Le, & McCloy, 2011), a more funda- mental issue with MTMM-based evaluations of AC functioning is their oversimplification of components of variance underlying assessor ratings.

For example, the major CFA-based summaries of components of variance underlying assessor ratings generally only describe variance in terms of dimension, exercise, and residual effects (e.g.,

Bowler & Woehr, 2006; Lievens & Conway, 2001), or general performance, exercise, and residual effects (e.g., Lance et al.,

2004). In contrast, past studies that have adopted alternative meth- ods that go beyond MTMM conceptualizations of AC functioning have revealed a much richer set of variance components underly- ing assessor ratings (e.g., Arthur, Woehr, & Maldegen, 2000; Bowler & Woehr, 2009; Lievens, 2001a, 2001b, 2002). In general, the oversimplification of sources of variance characteristic of MTMM-based AC research appears to be an unfortunate carryover from the application of the original Campbell and Fiske (1959) criteria to evaluations of AC functioning (Hoffman & Meade,

2012; Howard, 2008).

Another characteristic of MTMM-based AC research, which is an artifact of the oversimplification we have described, is its confounding of what would be viewed as reliable and unreliable sources of variance from the perspective of interrater reliability. This confounding stems from the fact that each of the PEDRs subjected to MTMM and CFA analyses typically reflect an aver- age of assessor ratings or the ratings of a single assessor (Lievens

& Conway, 2001; Sackett & Dreher, 1982). Distinguishing be- tween sources of consistency and inconsistency among different assessors’ ratings requires that more than one assessor provide a PEDR for each assessee on each dimension– exercise (D–E) unit and that those ratings are not averaged across assessors prior to analyses.1

**Reliable and Unreliable Variance in Assessor Ratings**

As noted recently by Putka and Sackett (2010), and long rec- ognized in the generalizability (G) theory literature, the composi- tion of reliable and unreliable variance depends on generalizations one wishes to make regarding the scores in question (Cronbach et al., 1972). Given that issues of reliability and its composition are complex, we first focus on a situation where one simply desires to generalize AC ratings from one or more assessors to other asses- sors. When estimating interrater reliability, reliable sources of variance are components of between-assessee variance in assessor ratings that are consistent across assessors, and unreliable sources of variance are components of between-assessee variance in as- sessor ratings that are inconsistent across assessors (LeBreton & Senter, 2007; Putka & Sackett, 2010; Schmidt & Hunter, 1989,

1996). Framed another way, reliable sources of variance contribute to similarity in the rank ordering of assessees based on ratings made by different assessors, and unreliable sources contribute to differences in rank ordering of assessees based on ratings made by different assessors.2 As noted previously, to distinguish between

1 A dimension-exercise (D–E) unit simply refers to the specific dimension– exercise combination for which postexercise dimension ratings (PEDRs) are gathered (e.g., ratings for Dimension 1–Exercise A, ratings for Dimension 2–Exercise A). Throughout this article, we refer to both D–E units and PEDRs.

2 Note that we are defining error here in relative, rather than absolute, terms (Brennan, 2001; Cronbach et al., 1972). If we were defining error in absolute terms, any effects that contribute to differences between assessors’ ratings—including those that do not contribute to between-assessee vari- ance (e.g., rater leniency/severity effects in a fully crossed design)—would contribute to error. Absolute error terms typically characterize indexes of interrater *agreement*, whereas relative error terms typically characterize indexes of interrater *reliability* (LeBreton & Senter, 2007; Putka & Sackett,

2010).

these sources of variance, one needs to adopt a measurement design in which more than one assessor rates each assessee on the variable of interest (e.g., a given D–E unit, a given dimension) and then analyze the resulting ratings.

Despite our initial focus on the interrater reliability of assessor ratings, we recognize that researchers may also be interested in generalizing ratings across other facets of their measurement design. For example, if exercises are designed to capture a similar aspect of dimensional performance, then one may also wish to assess how well ratings for a given dimension generalize across exercises (i.e., incon- sistency in an assessee’s performance across exercises would be viewed as error). In contrast, if exercises are designed to capture distinct aspects of performance on a dimension, then one may not be interested in generalizing ratings for that dimension across exercises, realizing that each exercise was meant to offer a different perspective. Ultimately, the measurement facet(s) one wishes to generalize AC scores across (e.g., assessors, exercises) is a decision to be made by individual researchers given their measurement situation (Cronbach et al., 1972; Putka & Sackett, 2010). Therefore, we revisit this issue when presenting our results and provide a concrete illustration of how the composition of reliable and unreliable variance can change de- pending on the generalizations one wishes to make regarding assessor ratings.

**Components of Variance Underlying Assessor Ratings**

Identifying components of variance that contribute to reliable and unreliable variance in assessor ratings requires not only being clear about defining reliable and unreliable variance but also carefully considering the measurement design underlying the rat- ings (e.g., crossed, nested, ill-structured; Putka & Sackett, 2011). Given the complexity of the issues involved, it is instructive to frame the ensuing discussion with a concrete example to avoid misinterpretation.

Imagine that 100 assessees complete four AC exercises, and each exercise comprises the same eight dimensions. Moreover, assume that two assessors independently observe and rate all assessees on each dimension after the completion of each exercise. That is, from a G-theory perspective, PEDRs are gathered using a fully crossed design (Cronbach et al., 1972). Though such a design is not very realistic, for pedagogical purposes such simplicity offers a useful point of departure—we introduce a more realistic design when we broach sources of unreliability in ratings. From the perspective of the resulting data set, we now have 64 columns of PEDRs (Two Assessors Four Exercises Eight Dimen- sions).

In the section that follows, we describe components of variance underlying assessors’ PEDRs in the example above and which ones contribute to reliable and unreliable variance from the per- spective of estimating interrater reliability. Though some research- ers have argued that PEDRs should not be the focus of the AC research given that they have traditionally been viewed as defi- cient, unreliable single-item indicators of a candidates’ AC per- formance, we believe this call is premature (e.g., Arthur et al.,

2008; Rupp et al., 2008; Thornton & Rupp, 2006). Examining sources of variance underlying PEDRs—which is the level at which most ratings in operational ACs appear to be gathered (Woehr & Arthur, 2003)—is essential for comprehensively parti- tioning assessee-, dimension-, exercise-, and assessor-related com-

ponents of variance. Specifically, aggregating PEDRs—which from a levels-of-analysis perspective reflect ratings of D–E units—to either the dimension level, exercise level, or overall AC level prior to variance partitioning confounds distinct sources of variance and therefore limits insights that can be gained into the composition of variance underlying assessor ratings.

Despite our initial focus on partitioning variance in PEDRs, we recognize that some level of aggregation of PEDRs typically occurs in practice to facilitate operational use of the AC data (e.g., overall assessment scores for selection and promotion-focused ACs, dimension-level scores for development-focused ACs; Kun- cel & Sackett, 2012; Spychalski, Quinones, Gaugler, & Pohley,

1997). Therefore, when providing our results, we illustrate how one can rescale variance components derived from analyzing PEDRs to estimate the expected composition of variance underly- ing dimension-level scores, exercise-level scores, and overall-AC scores formed from aggregation of PEDRs. In other words, we show how decomposing variance in raw PEDRs offers flexibility for estimating the composition of variance and functioning of AC scores at not only the D–E unit level (raw PEDRs) but also higher levels of aggregation.

**Sources of Reliable Variance in Assessor Ratings**

Given the measurement design outlined in the example above and the definition of interrater reliability offered earlier, there are four sources of reliable variance in assessors’ PEDRs that can be uniquely estimated: (a) assessee main effect variance, (b) Assessee Dimen- sion interaction effect variance, (c) Assessee Exercise interaction effect variance, and (d) Assessee Dimension Exercise interaction effect variance.3 Table 1 provides definitions of each of these com- ponents and assumes that the components are estimated using random effects models. In the sections that follow, we briefly describe the theoretical underpinnings of each component and highlight the limi- tations that MTMM-based AC research has for isolating the contri- bution of these sources of variance to assessor ratings.

**Assessee main effect variance.** Assessee main effect variance can loosely be thought of as reflecting a reliable general factor in assessor ratings.4 The presence of a general factor has been given scant attention in the AC literature and usually goes unmodeled in common CFA models of PEDRs (e.g., correlated dimension– correlated exercise [CD–CE], correlated dimension– correlated uniqueness models [CD–CU]). Nevertheless, several recent studies in the AC literature have found evidence of a general factor underlying assessor ratings (e.g., Hoffman et al., 2011; Lance et al., 2000, 2004; Lance, Foster, Nemeth, Gentry, & Drollinger,

2007). These findings are consistent with job performance litera- ture that continues to suggest the presence of a general factor underlying the performance domain—from which the tasks that

3 Note that dimension main effects, exercise main effects, and Dimen- sion Exercise interaction effects do not contribute to observed between- assessee variance because their underlying effects are constants across assessees—thus contrary to past claims, they do not contribute to either reliable or unreliable variance in assessor ratings (Bowler & Woehr, 2009).

4 Technically, there is nothing within the definition of random effects models that underlie G theory that necessitate assessee main effect variance being viewed as reflecting only a single homogeneous general factor. Nonetheless, the analogy offered here is useful for contrasting this variance component with the other components of reliable variance described.

Table 1

*Decomposition of Observed Variance in Assessor Postexercise Dimension Ratings*

Variance component Substantive meaning Covariance interpretation

Reliable variance Regarding assessees’ performance on a given dimension–exercise unit, this component implies that . . .

2 some assessees perform better than others,

assessee

Expected level of covariance between *two different assessors’ ratings* of the *same* dimension–exercise unit that is . . .

neither dimension- nor exercise-specific (i.e.,

regardless of dimension, exercise, or assessor

2 some assessees perform better on some

assessee dim

akin a general performance factor)

specific to the dimension examined

dimensions than others, regardless of exercise or assessor

2 some assessees perform better on some

assessee ex

specific to the exercise examined

exercises than others, regardless of dimension or assessor

2 some assessees perform better on some

assessee dim ex

specific to the dimension–exercise combination

dimension–exercise combinations than others, regardless of assessor

Unreliable variance Regarding assessees’ performance on a given dimension–exercise unit, this component implies that . . .

2 some assessees are rated higher by some

assessee assessor

examined

Expected level of covariance between the same assessor’s ratings of two *different* dimension–exercise units that is . . .

neither dimension- nor exercise-specific, but is

assessors than others, regardless of dimension or exercise

2 some assessees are rated higher by some

assessee dim assessor

specific to the assessor making the rating

(akin to general rater halo)

specific to the dimension examined and

assessors than others—but it depends on the dimension

2 some assessees are rated higher by some

assessee ex assessor

assessor making the rating (akin to dimension-specific rater halo)

specific to the exercise examined and assessor

assessors than others—but it depends on the exercise

2 some assessees are rated higher by some

4-way interaction residual

making the rating (akin to exercise-specific rater halo)

specific to the dimension-exercise combination

assessors than others—but it depends on the dimension-exercise combination

2 some assessors tend to give more lenient/

assessor

examined and assessor making the rating

severe ratings than others, regardless of dimension or exercise

2 some assessors tend to give more lenient/

assessor dim

severe ratings than others—but it depends on the dimension

2 some assessors tend to give more lenient/

assessor ex

severe ratings than others—but it depends on the exercise

2 some assessors tend to give more lenient/

assessor dim ex

severe ratings than others—but it depends on the dimension-exercise combination

*Note*. The assignment of components to reliable and unreliable variance in this table assumes that one is interested in generalizing ratings across assessors only. As we note later, which components contribute to reliable and unreliable variance will shift if one desires to generalize ratings across other facets of one’s measurement design (e.g., exercises). Dim dimension; ex exercise.

Component only contributes to between-assessee variance and unreliable variance when assessees and assessors are not fully crossed. Because the contribution of these effects depends on a lack of full crossing between assessees and assessors, it is difficult to provide clear “covariance” based interpretation of them (see also Footnote 7).

compose an AC presumably sample (e.g., Hoffman, Lance, By- num, & Gentry, 2010; Viswesvaran, Schmidt, & Ones, 2005).

Despite the consistency with the job performance literature, emerging evidence for a general factor in assessor ratings does not clarify the magnitude of assessee main effect variance because such evidence has largely been based on studies in which PEDRs that had been aggregated across assessors were modeled. Thus, the general factor from such studies reflects not only assessee main effect variance but also, to an unknown extent, a source of unreliable variance, namely, Assessee Assessor variance (see Table 1).5

**Assessee**  **Dimension interaction effect variance.** In the context of MTMM CFA models of PEDRs, this variance compo-

nent would be akin (but not identical) to variance in PEDRs attributable to dimension factors that are modeled as uncorrelated with each other and other factors in the model (Woehr, Putka, & Bowler, 2012). Though one can gain a rough sense of an upper bound estimate for the magnitude of Assessee Dimension vari- ance by examining the magnitude of “dimension effects” from past CFA research (e.g., Bowler & Woehr, 2006; Lievens & Conway,

5 Note that if assessors are not fully crossed with assessees (which they are often not in operational AC data), then any variance attributable to assessor main effects (e.g., assessor leniency/severity differences) will also be reflected in general factor variance estimated by CFA models of aggregated or single-assessor PEDRs.

2001), such an estimate would be highly contaminated. For example, given that MTMM CFA models have largely been fitted to aggregated or single-assessor PEDRs, variance attributable to dimension factors not only reflects Assessee Dimension variance but also a source of unreliability, namely, Assessee Dimension Assessor variance (again, see Table 1). Moreover because dimensions are often modeled as correlated in MTMM CFA models and no general performance factor is typically specified (e.g., consider CD–CE, CD–CU models), variance attributable to dimension factors in these models is not purely a function of Assessee Dimension effects but also reflects unmodeled general factor variance and other sources of covariance among dimensions.

Despite the presence of these additional sources of variance, summaries of variance in PEDRs often suggest that variance attributable to dimension factors is often small to moderate (e.g., Bowler & Woehr, 2006; Connelly, Ones, Ramesh, & Goff, 2008; Lance et al., 2004; Sackett & Dreher, 1982). As such, we expect that Assessee Dimension variance would be even smaller than what has been attributed to dimension factors in past summaries of AC research due to the confounding issues we have noted.

**Assessee**  **Exercise interaction effect variance.** In the con- text of MTMM CFA models of PEDRs, this variance component would be akin (but not identical) to variance in PEDRs attributable to exercise factors that are modeled as uncorrelated with each other and other factors in the model (Woehr et al., 2012). As with the other sources of reliable variance described, one can only get a crude sense of the magnitude of Assessee Exercise interaction effect variance based on past CFA research—again being wary of confounds. For example, given that MTMM CFA models have been fitted largely to aggregated or single-assessor PEDRs, vari- ance attributable to exercise factors not only reflects Assessee Exercise variance but also unreliable Assessee Exercise Assessor variance (see Table 1). Moreover because exercises are modeled as correlated in CD–CE models, variance attributable to exercise factors in such models is not purely a function of Assessee Exercise effects but also reflects variance that is shared across exer- cises (e.g., an unmodeled general factor or other source of covariance among exercises).

Perhaps as a result of these additional sources of variance, MTMM CFA models of PEDRs often suggest that variance attrib- utable to exercise factors is typically sizable (Bowler & Woehr,

2006; Connelly et al., 2008; Lance et al., 2004; Lievens & Con- way, 2001; Sackett & Dreher, 1982). However, as with the previ- ously mentioned components of reliable variance, it is difficult to estimate how large the Assessee Exercise component actually is because it represents only a portion of what has typically been interpreted as exercise factor variance.

**Assessee**  **Dimension**  **Exercise interaction effect variance.** This final source of reliable variance in PEDRs has received little or no direct attention in AC research. In the context of traditional MTMM CFA models of PEDRs, this component is confounded with residual error (uniqueness). Despite the lack of empirical attention given to this component, there are theoretical reasons to believe that Assessee Dimension Exercise effects may be an important component of reliable variance in assessor ratings. For example, the notion that a dimension may manifest itself differently across exercises is consistent with trait activation theory (TAT) and opens up the possibility that assessees may legitimately perform differently on a dimension depending on

exercise used to elicit it (Borman, 2012; Hoffman et al., 2011; Howard, 2008; Lievens et al., 2006; Melchers & König, 2008). That is, from the perspective of the variance components dis- cussed, ACs may capture both reliable, situationally stable dimen- sion effects (e.g., Assessee Dimension effects) *and* reliable, situationally varying dimension-related effects (e.g., Assessee Dimension Exercise effects). Indeed, in cases where exercises are designed to measure different aspects of dimensional perfor- mance, failure to find notable Assessee Dimension Exercise effects would be undesirable as it might signify (a) unnecessary redundancy of measurement or (b) potential deficiency in con- struct coverage (Borman, 2012; Howard, 2008; Lievens et al.,

2006).

**Sources of Unreliable Variance in Assessor Ratings**

From the perspective of interrater reliability, each of the four sources of reliable variance discussed above has an unreliable analogue that reflects variance that is idiosyncratic to individual assessors (i.e., the unreliable components simply reflect incon- sistency of the given effect across assessors; see Table 1). For example, the unreliable analogue of assessee main effect vari- ance is assessee assessor interaction effect variance. In more general terms, these components can be thought of as different varieties of rater-specific factor error or idiosyncratic rater halo (Le & Putka, 2006; Viswesvaran et al., 2005). For example, the Assessee Dimension Assessor interaction effect can be interpreted as a dimension-specific assessor halo effect, in that it reflects shared variance between the vectors of a given assessor’s PEDRs that is specific to the dimension and assessor considered.

Though the AC literature has been relatively silent on com- ponents of unreliable variance outside lab settings (e.g., Lievens, 2001a, 2001b), the general literature on observer rat- ings suggests that these components may be fairly small in the context of PEDRs. For example, Hoyt and Kerns (1999) meta- analyzed variance components underlying observer ratings fo- cusing specifically on Ratee Rater interactions effects (akin to the Assessee Assessor effects) and rater main effects. Under conditions that resembled those that arguably character- ize a carefully implemented AC, namely (a) directly observable behaviors explicitly tied to scale anchors, (b) high levels of rater training (e.g., 25 hr), and (c) having raters observe and rate ratees’ behavior on the same occasion, Hoyt and Kerns found that the Ratee Rater interaction variance only ac- counted for an average of 3%, 6%, and 9% of observed variance in ratings made by different raters for each ratee, respectively.6

Furthermore, although limited, past AC research suggests that assessors can be quite reliable and accurate in their ratings (e.g., Arthur & Day, 2011; Connelly et al., 2008; Lievens, 2001a,

6 Note that if Hoyt and Kerns (1999) standardized their variance com- ponent estimates assuming a fully crossed design (i.e., one in which the same raters rated all ratees), then these percentages would have been somewhat higher given that observed variance would not have included rater main effects (i.e., the divisor for calculating these percentages would have been lower). However, given that most operational AC ratings are unlikely to be based on fully crossed designs, Hoyt and Kerns’ (1999) decision to standardize their estimates based on a completely nested design likely makes them closer to what one would see in AC practice.

2001b; Tsacoumis, 2007). Thus, we expect the set of Assessee Assessor-related interaction effects (e.g., Assessee Assessor, As- sessee Dimension Assessor, Assessee Exercise Assessor) to be far smaller than the set of reliable components previously discussed and fairly small in most carefully constructed and im- plemented ACs.

**Revisiting the Role of Measurement Design**

The decomposition of variance outlined in the previous sec- tions paints an unrealistically simple picture of sources of observed variance in assessor ratings. As evidenced by the G-theory literature, measurement design can have profound implications in terms of what sources of variance contribute to observed variance and error in scores (Brennan, 2001; Cron- bach et al., 1972). For example, if assessors were not fully crossed with assessees in our working example, then assessor main effects (often attributed to differences in rater leniency/ severity; Hoyt, 2000; Hoyt & Kerns, 1999; Putka, Le, McCloy,

& Diaz, 2008; Viswesvaran et al., 2005), Assessor Dimen- sion, Assessor Exercise, and Assessor Dimension Exercise interaction effects would also contribute to unreliable variance (see Table 1).7 As such, measurement design not only has implications for the magnitude of unreliable variance (i.e., larger in designs that are not fully crossed if variance attribut- able to the aforementioned effects is nonzero) but also has clear implications for how one describes the composition of observed variance in assessor ratings. In large operational ACs, assessors are rarely fully crossed with assessees, and thus, assessor main effects and the assessor-related interactions noted earlier have the potential contribute to unreliability in assessor ratings (Lievens, 2009; Putka et al., 2008; Spychalski et al., 1997).

**Moving Beyond MTMM Conceptualizations of AC Functioning**

In light of limitations with MTMM-based approaches to exam- ining AC functioning, researchers have begun to use random effects models that can provide richer decompositions of variance in assessor ratings and address the confounds described above (e.g., Arthur et al., 2000; Bowler & Woehr, 2009; Lievens, 2001a,

2001b, 2002). Specifically, the example studies noted above share four key characteristics that are critical for disentangling the com- ponents of reliable and unreliable variance described. Namely, they all (a) explicitly included assessor-related effects in their models, (b) obtained PEDRs for any given D–E unit from more than one assessor per assessee, (c) analyzed disaggregated PEDRs (i.e., not averaged assessors), and (d) specified models that parti- tioned PEDRs into far more components than the typical MTMM- based study focused on issues of construct validity.8 Though random effects models can potentially help eliminate the cofounds characteristic of MTMM- and CFA-based approaches previously noted, the models are assumed to have been properly specified and applied to data that allow meaningful generalization of parameter estimates. Unfortunately, as we describe next, these studies pro- vide limited perspectives on components of reliable and unreliable variance in assessor ratings because of (a) model misspecification issues, (b) limitations of the data examined, and (c) suboptimal use of available data.

**Model Misspecification Issues**

There are multiple sources of misspecification in the random effects models used in the Arthur et al. (2000) and Bowler and Woehr (2009) studies that limit their ability to produce mean- ingful variance component estimates. One source stems from the fact they did not specify all possible interactions among assessees, assessors, dimensions, and exercises in their models. By omitting several interaction terms, these studies confound estimates of residual error variance with variance attributable to unmodeled interactions (including a source of reliable vari- ance—Assessee Dimension Exercise effects) and likely inadvertently violate a key assumption underlying the models examined, namely, independence of residuals (Searle, Casella,

& McCulloch, 2006). Specifically, to the extent that any vari- ance in assessor ratings is attributable to omitted interactions, there would be a source of nonindependence among residuals attributable to such interactions (e.g., residuals that share an assessee– dimension– exercise combination would not be inde- pendent). Failing to account for such nonindependence has been found to bias variance component estimates both upwardly and downwardly for other terms included in the model (e.g., Bost,

1995; Kenny & Judd, 1986; Maxwell, 1968; Putka et al., 2008; Smith & Luecht, 1992).

Another source of model misspecification in the Arthur et al. (2000) and Bowler and Woehr (2009) studies arises from dis- connects between the structure of the models fitted to the data and the structure of the underlying data. Specifically, Arthur et al. (2000) modeled assessees as being nested within assessors when in reality they were not because each assessee was rated by between three and four assessors. If assessees were nested within assessors, then by definition, each assessee would only be rated by a single assessor (Brennan, 2001; Putka et al., 2008; Searle et al., 2006). By specifying assessees as nested within assessors, Arthur et al. (2000) unnecessarily confounded as- sessee main effects (a source of reliable variance) with As- sessee Assessor interaction effects—a source of unreliable variance (Hoyt, 2000; Hoyt & Kerns, 1999; Putka et al., 2008;

7 Note that if assessors are not fully crossed with assessees *and* if the effects noted here are not explicitly specified in one’s variance decompo- sition model (regardless of whether that model is CFA- or random effects- based), these effects will manifest themselves as other sources of variance noted earlier (i.e., there will be confounding of distinct sources of vari- ance). Namely, variance attributable to assessor main effects will manifest as Assessee Assessor interaction variance, Assessor Dimension effects will manifest as Assessee Dimension Assessor variance, Assessor Exercise effects will manifest as Assessee Exercise Assessor variance, and Assessor Dimension Exercise effects will manifest as residual variance.

8 Note that Woehr et al. (2012) recently illustrated how random effects models can be directly related to Campbell & Fiske’s (1959) MTMM framework. However, as past studies have illustrated (e.g., Arthur et al.,

2000; Bowler & Woehr, 2009; Lievens, 2001a, 2001b, 2002), random effects model are quite flexible and can expand to account for a much richer set of variance components than the ones examined in Woehr et al. (2012).

Schmidt & Hunter, 1996).9 Consequently, the true score (or universe score) component of any reliability (generalizability) coefficient formed on the basis of the components reported by Arthur et al. (2000) would reflect a key source of unreliable variance to an unknown extent.

In Bowler and Woehr (2009), the disconnect between the fitted model and underlying data stems from how assessors were coded for purposes of fitting the model. Specifically, Bowler and Woehr (2009) coded the two assessors who rated each assessee on a given D–E unit as “1” and “2,” as opposed to giving each of the 35 assessors who participated in the study a unique assessor identifi- cation code. Unfortunately, such coding makes it impossible to meaningfully estimate or interpret assessor main effects and all assessor-related interaction effects (Putka et al., 2008, 2011). For example, Assessor 1 may actually reflect two *different* assessors across assessees, whereas Assessors 1 and 2 may actually reflect the *same* assessor across assessees. Couple this with the fact that Bowler and Woehr (2009) modeled assessor ratings that were made *after* assessors discussed the performance of the assessees (and thus did not represent the independent perspectives of asses- sors), and it helps explain why they found no variance in ratings attributable to assessor main effects and assessor-related interac- tions.

**Limitations of the Data**

Beyond the field studies discussed above, Lievens (2001a,

2001b, 2002) published a series of lab studies that estimated variance in assessor ratings attributable to various combinations of assessee, assessor, dimension, and exercise-related effects using random effects models. These studies largely avoid the misspeci- fication issues noted previously and suggest that assessors can potentially serve as accurate judges of assessee behavior in the context of ACs. Unfortunately, these studies are of limited use for informing the composition of reliable and unreliable variance in operational assessor ratings due to the data on which they are based.

For example, Lievens’ (2002) lab study provides insight into components of *within-assessee* variance in PEDRs (across dimen- sions, exercises, and assessors) but does not allow one to evaluate reliable and unreliable sources of variance in PEDRs because it did not partition *between-assessee* variance. All methods of estimating reliability, regardless of whether they are based on classical test theory, G-theory, or CFA traditions, require partitioning of *between-person* variance into true (universe) score and error com- ponents (Cronbach et al., 1972; Putka & Sackett, 2010; Schmidt & Hunter, 1989). Though Lievens’ (2001a, 2001b) studies did par- tition between-assessee variance, the variance components reflect- ing assessee main effects and assessee-related interactions were estimated on the basis of only *four* simulated assessees within the context of a lab study. As such, the results arguably provide a limited perspective on the expected contribution of such effects to *between-assessee* variance in operational AC data.

**Suboptimal Use of Available Data**

Another limitation of three of the four random effects model studies we have noted (i.e., Arthur et al., 2000; Lievens, 2001b,

2002) stems from suboptimal use of the available data. Specifi-

cally, these studies unnecessarily discarded nontrivial portions of their data to achieve a balanced design operating under the false assumption that such a design was needed to estimate variance components. For example, Arthur et al. (2000) discarded one of four exercises and five of nine dimensions to achieve a balanced design. Similarly, Lievens (2002) randomly discarded six of 26 psychologist assessors and seven of 27 student assessors to achieve a balanced design. Discarding of data in this manner is suboptimal in that not all of the data were used, and therefore, variance component estimates are less stable than they could have been had all data been used (Enders, 2010). Additionally, discarding data introduces an unaccounted-for source of variation into the esti- mates—namely, variation reflecting the sampling of data chosen for inclusion in the analysis (i.e., starting with the same data set, one could obtain different results with a different subsampling of assessors, despite using the same assessees; Putka et al., 2011).

Though balanced designs are needed to estimate variance com- ponents using the analysis-of-variance- (ANOVA)-based estima- tors (specifically, expected mean square, or estimators) discussed by Cronbach et al. (1972) in their seminal treatise on G theory, such ANOVA-based estimators have generally fallen out of favor in the broader literature on variance component estimation (Searle et al., 2006). Indeed, maximum-likelihood-based estimators— which do not require one to have a balanced design and have been available to researchers for decades in common statistical pack- ages such as SPSS and SAS—appear to be favored (DeShon, 1995; Greguras & Robie, 1998; Marcoulides, 1990; Putka et al., 2008; Putka & Sackett, 2010; Searle et al., 2006).10

**Summary**

The previous sections reveal that unconfounded estimates of components of reliable and unreliable variance in assessor ratings have yet to be provided by MTMM-based and random effects model– based AC research. In the current study, we were able to obtain data and fit models that provide unique estimates for all components summarized in Table 1 and clarify the contribution that assessee-, dimension-, exercise-, and assessor-related effects make to assessor ratings.

9 Given the nature of their measurement design, Arthur et al. (2000) could have modeled assessees and assessors as crossed random factors and used modern variance component estimation methods to deal with the lack of complete crossing of assessees and assessors (Searle et al., 2006). This would have allowed Arthur and colleagues to uniquely estimate variance attributable to assessee main effects, assessor main effects, and Assessee Assessor interaction effects and completely avoid the issues raised here (e.g., Putka et al., 2008). In Arthur et al.’s defense, the computational demands of such an analysis would have been quite steep, and as such, perhaps not possible given the computational resources Arthur and col- leagues had access to at the time of their study (e.g., Bell, 1985).

10 Despite their benefits and favor among statisticians who actively study variance components (e.g., Searle et al., 2006), maximum likelihood variance component estimators do have some potential limitations (e.g., high memory requirements, estimation issues when interactions involving fixed factors are modeled, and normality assumptions).

**Overview of Samples**

**Method**

**AC Development**

To help maintain test security, we developed new exercises and rating scales prior to each year’s AC based on a common set of job

Data for this study come from three independent samples of candidates from three separate implementations of an operational AC conducted between 2002 and 2008. This high-stakes AC was designed to evaluate internal candidates for promotion to first-line supervisory positions at a large government organization. Sample

1 comprised the 153 unique candidates from the 2002 AC imple- mentation, Sample 2 the 198 unique candidates from the 2006 AC, and Sample 3 the 282 unique candidates from the 2008 AC. In other words, there was no overlap among candidates in the samples examined. Table 2 provides demographic characteristics of these candidates.

**Assessment Center Design**

Despite a few differences that are detailed below, the general process of administering the ACs was the same each year. The assessment process involved administration of three or four exer- cises, each designed to measure three to 10 dimensions (see the Appendix). We followed recommendations regarding being trans- parent about the dimensions assessed, providing candidates with a listing of the dimensions as part of the instructions given with each exercise (Arthur & Day, 2011; Lievens, 1998).11

Assessors were experienced supervisors and managers from the participating organization. Twenty-seven assessors participated in the first AC, 33 in the second, and 46 in the third. For each candidate, two assessors provided PEDRs at the conclusion of each exercise resulting in a total of between six (Sample 3) and eight unique assessors (Samples 1 and 2) per candidate (i.e., a candidate was rated by a different pair of assessors for each exercise). The measurement design linking candidates to assessors could be de- scribed as ill structured in that candidates were not fully crossed with assessors, nor were candidates or assessors nested within one another (Putka et al., 2008). Because assessors and candidates were both from the same organization, assessors were randomly matched to candidates by a computer algorithm that minimized the potential for familiarity conflicts between assessors– candidate pairs based on familiarity ratings provided by assessors in advance of each AC.

Table 2

*Demographic Characteristics of Assessment Center Candidates*

Sample

analysis data. Each year’s exercises and their associated rating scales were developed over the series of four 2-day workshops led by a team of experienced industrial and organizational (I–O) psychologists and pilot tested among new incumbents in the target job prior to their operational implementation. Development of the exercises followed a rigorous content-oriented process that began with (a) reviewing critical task lists (identified via a comprehen- sive job analysis of the target job), (b) discussing which tasks were amenable to be simulated within the context of the AC, and (c) selecting subsets of critical tasks to target for simulation within each exercise. When selecting subsets of tasks, the I–O psychol- ogist project leads reviewed the knowledges, skills, and abilities (KSAs) required for performing those tasks (based on the job analysis data) with an eye toward maximizing the breadth and depth of dimensional assessment. As each exercise was developed, certain dimensions that were initially targeted for assessment by the exercise were dropped from consideration if they failed to offer sufficient opportunities for candidates to display behaviors related to that dimension (Howard, 2008; Lievens et al., 2006).

**AC Implementation**

Assessors used behavioral summary scales custom developed for each D - E unit when making their PEDRs. The scales were comparable in structure to the example rating scale provided by Tsacoumis (2007), and the multiple behavioral anchors that com- prised them could be tied back to multiple, job critical KSAs identified in the job analysis. Assessors provided both independent prediscussion and postdiscussion PEDRs. Given the focus of this study, we analyzed assessors’ prediscussion ratings to evaluate the composition of variance in ratings based on the independent view- points of the assessors.

Prior to each AC, assessors received 32 hr of training; key components included (a) orienting assessors to the AC process, (b) reviewing exercises and rating materials, (c) reviewing and allow- ing assessors to practice the roles they needed to play (e.g., behavioral observer, note taker, behavioral categorizer, role player, rater), (d) instructing assessors on how to document and categorize candidate behavior into dimensions, (e) raising assessors’ aware- ness regarding common types of rating errors (e.g., leniency, severity, central tendency, halo), (f) having assessors indepen- dently record the behavior of and evaluate mock candidates using

Sex

Variable

1 2 3

the rating scales, and (g) having a follow-up group discussion to calibrate assessors’ use of the rating scales and instill confidence in

Male 139 180 257

Female 14 18 25

Race/ethnicity

White 106 147 220

African American 21 27 26

Hispanic 19 14 27

Other 7 10 8

*Note*. The mean organizational tenure of candidates in the samples were as follows: Sample 1: 13.46 years (*SD*  3.11), Sample 2: 10.75 years (*SD*  5.36), and Sample 3: 11.12 years (*SD*  5.19).

the assessment process.

11 Though being transparent about dimensions may seem unusual given the high-stakes nature of the assessment, the sponsoring organization viewed transparency as issue of fairness to candidates and believed that candidates should be provided the names of the dimensions on which they were being evaluated for promotion, as well as the specific KSAs from the job analysis on which they were based.

**Analyses**

We first estimated two reliability coefficients for PEDRs in each sample. The first is an estimate of single-rater reliability that reflects the expected proportion of between-candidate variance in PEDRs based on a single assessor’s initial ratings that is attribut- able to reliable score differences between candidates. The second is an estimate of interrater reliability for PEDRs formed from averaging two assessors’ initial ratings. This latter estimate reflects the expected proportion of between-candidate variance in average PEDRs that is attributable to reliable score differences between candidates. Given the ill-structured nature of the measurement design, both reliability estimates were based on Putka et al.’s (2008) *G*(*q,k*) coefficient.

Next, we decomposed the variance in assessors’ PEDRs by fitting a linear random effects model to assessors’ initial ratings that (a) imposed a variance components covariance structure on the random effects design matrix and (b) specified candidates, assessors, dimensions, and exercises as crossed random factors (Littell, Milliken, Stroup, & Wolfinger, 1996; Searle et al., 2006). As noted by Putka et al. (2008), when confronted with ill- structured ratings data, it is possible to fit a model in which factors are specified as crossed (despite the lack of full crossing among the factors) and capitalize on modern variance component estimation techniques that are robust to missing data to recover the parameters of interest (DeShon, 1995; Marcoulides, 1990; Searle et al., 2006). To estimate variance components underlying this model, we used the HPMIXED procedure in SAS, which is a variant on SAS’s MIXED procedure specifically designed for estimating variance components in large, sparse data structures (SAS Institute,

2011).12 Sparseness in the current study arises from assessors not

being fully crossed with candidates.

Last, we rescaled variance components from the analyses using rules outlined in the G-theory literature to estimate the contribution of each component to variance in aggregate dimension-level scores (i.e., average of PEDRs across exercises for a given dimension), exercise-level scores (i.e., average of PEDRs across dimensions within a given exercise), and overall-AC scores (i.e., average of PEDRs across dimensions and exercises; Shavelson & Webb,

1991). We then used these results to provide expected levels of AC score reliability as a function of the (a) level of AC score examined (e.g., D–E unit-level, dimension-level) and (b) generalizations we desired to make regarding those scores (e.g., generalizing across assessors only, or generalizing dimension scores across assessors and exercises).

**Results**

The average single-rater reliability for PEDRs across all three samples was .76 (*SD* across D–E units: .05; range: .59 to .87), and the average interrater reliability (based on the average of two assessors) was .87 (*SD* across D–E units: .03; range across D–E units: .75 to .93).13 These results suggest that the vast majority of variance in PEDRs, whether based on the ratings of a single assessor per candidate, or the average of two assessors per candi- date, reflects reliable variance when viewed from the perspective of interrater reliability. This runs somewhat contrary to claims in the AC literature that PEDRs are akin to unreliable item-level units of information (e.g., Arthur et al., 2008; Rupp et al., 2008), but is

quite consistent with our expectations based on Hoyt and Kerns’ (1999) more general work on observer ratings.

**Composition of Reliable and Unreliable Variance in PEDRs**

Table 3 shows the decomposition of reliable and unreliable variance in PEDRs based on ratings of a single assessor per candidate (not the same assessor per candidate). The components in the table are organized into three sets: (a) sources of reliable variance in assessor ratings, (b) sources of unreliable variance in assessor ratings, and (c) other components that do not contribute to between-candidate variance. For each sample, three columns of information are provided. The first shows the percentage of ex- pected total variance attributable to each component. These reflect the raw variance components estimated by SAS, divided by the sum of all components for the sample, and multiplied by 100. Though such percentages have often been reported in past AC studies with random effects models (e.g., Arthur et al., 2000; Bowler & Woehr, 2009; Lievens, 2001a), the concept of expected total variance does not carry much practical value in that some of its components do not contribute to observed between-candidate variance (Brennan, 2001; Cronbach et al., 1972; Cronbach, Lynn, Brennan, & Haertel, 1997; Woehr et al., 2012). Therefore, the second column under each sample shows the percentage of ex- pected between-candidate variance attributable to each component. Each component of between-candidate variance is in turn either a source of reliable variance or unreliable variance in assessor rat- ings. Thus, the third column under each sample shows the per- centage of reliable or unreliable variance attributable to each variance component.

All three samples reveal a similar pattern of findings for reliable sources of variance. Specifically, variance attributable to Assessee Exercise interaction effects accounted for the largest percentage of reliable variance in each sample (42.8%–52.0%). In contrast, Assessee Dimension interaction effects accounted for the smallest percentage of reliable variance across the three sam- ples—in fact, a much smaller percentage than any other compo- nent of reliable variance (0.5%–1.8%) and smaller than variance typically attributable to dimension effects in past summaries of AC functioning (e.g., Bowler & Woehr, 2006; Lance et al., 2004; Lievens & Conway, 2001).

Though the general pattern of these findings may be viewed as somewhat expected based on past summaries of the literature, perhaps the most critical finding is that two other components that have received limited attention accounted for nearly half of the reliable variance in PEDRs. Specifically, the second largest source of reliable variance was attributable to Assessee Dimension Exercise interaction effects (28.6%–31.3%), and the third largest source of reliable variance was always assessee main effects (18.9%–24.4%). The magnitude of these latter two components clearly suggests that sources of consistency in different assessors’ PEDRs are a function of far more than traditional dimension and exercise effects.

12 The SAS code for fitting these models is provided in supplemental online material accompanying this article.

13 Specific interrater reliability estimates for each D–E unit are provided in the supplemental online material accompanying this article.

Table 3

*Variance Component Estimates for Postexercise Dimension Ratings by Sample*

Unreliable variance

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Sample 1 (%) |  |  |  | Sample 2 (%) |  |  |  | Sample 3 (%) |  |
| Between- | Between- |  |  | Between- | Between- |  |  | Between- | Between- |
|  | Total | candidate | candidate |  | Total | candidate | candidate |  | Total | candidate | candidate |
| Variance component | variance | variance | subtotal |  | variance | variance | subtotal |  | variance | variance | subtotal |
| Reliable variance |  |  |  |  |  |  |  |  |  |  |  |
| 2 candidate | 16.7 | 18.8 | 24.4 |  | 13.1 | 14.6 | 18.9 |  | 13.8 | 18.2 | 23.9 |
| 2  candidate dim | 1.2 | 1.4 | 1.8 |  | 0.3 | 0.4 | 0.5 |  | 1.0 | 1.4 | 1.8 |
| 2  candidate ex | 29.2 | 32.9 | 42.8 |  | 36.2 | 40.3 | 52.0 |  | 24.8 | 32.8 | 43.0 |
| 2  candidate dim ex | 21.2 | 23.9 | 31.0 |  | 19.9 | 22.2 | 28.6 |  | 18.0 | 23.9 | 31.3 |
| Subtotal | 68.3 | 77.0 |  |  | 69.6 | 77.5 |  |  | 57.7 | 76.3 |  |
| 2  candidate assessor | 0.0 | 0.0 | 0.0 |  | 2.9 | 3.2 | 14.1 |  | 0.0 | 0.0 | 0.0 |
| 2  candidate dim assessor | 8.6 | 9.7 | 42.3 |  | 0.0 | 0.0 | 0.0 |  | 6.5 | 8.5 | 36.0 |
| 2  candidate ex assessor | 2.0 | 2.3 | 10.0 |  | 0.0 | 0.0 | 0.0 |  | 2.7 | 3.5 | 14.8 |
| 2  4-way interaction residual | 8.5 | 9.6 | 42.0 |  | 15.5 | 17.3 | 76.7 |  | 6.9 | 9.1 | 38.2 |
| 2  assessor | 0.4 | 0.4 | 1.9 |  | 0.8 | 0.8 | 3.7 |  | 0.3 | 0.4 | 1.7 |
| 2  assessor dim | 0.4 | 0.5 | 2.0 |  | 0.5 | 0.5 | 2.4 |  | 0.8 | 1.1 | 4.5 |
| 2  assessor ex | 0.0 | 0.1 | 0.2 |  | 0.0 | 0.0 | 0.2 |  | 0.4 | 0.5 | 2.0 |
| 2  assessor dim ex | 0.4 | 0.4 | 1.7 |  | 0.6 | 0.6 | 2.9 |  | 0.5 | 0.6 | 2.7 |
| Subtotal | 20.4 | 23.0 |  |  | 20.3 | 22.5 |  |  | 18.0 | 23.7 |  |
| 2 dim | 5.0 | — | — |  | 8.5 | — | — |  | 6.1 | — | — |
| 2 ex | 4.6 | — | — |  | 0.0 | — | — |  | 13.8 | — | — |
| 2  dim ex | 1.7 | — | — |  | 1.7 | — | — |  | 4.4 | — | — |
| Estimated *G* (*q*,1) |  | 0.77 |  |  |  | 0.77 |  |  |  | 0.76 |  |

Other components

*Note*. Sample 1 *N*  153. Sample 2 *N*  198. Sample 3 *N*  282. Estimated *G* (*q*,1) reliable variance subtotal/(reliable variance subtotal unreliable variance subtotal) expected single-rater reliability for any given postexercise dimension ratings. Dashes indicate component not estimable. Dim dimension; ex exercise.

These variance components contribute to unreliable variance because candidates are not fully crossed with assessors. Given the ill-structured nature of the measurement design, their contributions to between-candidate variance have been rescaled using the *q*-multiplier underlying the single-rater reliability estimates.

With regard to sources of *unreliable* variance underlying asses- sor ratings, findings were less consistent across samples. We attribute this inconsistency in part to the sparseness of the mea- surement design underlying the data examined. Nevertheless, as a whole, the percentage of between-candidate variance in PEDRs attributable to unreliable variance was relatively small and stable across samples (22.5%–23.7%). In light of these observations, we decided not to focus on patterns of unreliable variance.

A final key observation regarding results presented in Table 3 is that nonresidual components consistently accounted for the vast majority of between-candidate variance in PEDRs (82.7%–

90.9%). This finding speaks to the comprehensiveness of the fully specified random effects model examined here in terms of ac- counting for variance in PEDRs as a function of substantively meaningful (nonresidual) effects.

**Moving Beyond the Interrater Reliability of PEDRs**

As noted earlier, variance components estimated for PEDRs can be rescaled to estimate the composition of reliable and unreliable variance underlying aggregate dimension-level, exercise-level, and overall-AC scores. Table 4 provides results of rescaling sample- size weighted averages of raw variance components outputted by SAS. The percentages reported under the dimension-level scores, exercise-level scores, and overall-AC score columns in Table 4 show the expected percentage of between-candidate variance at- tributable to each component, as well as the formulae used for

rescaling the raw variance components based on the G-theory literature (e.g., Shavelson & Webb, 1991, pp. 86 – 89). The for- mulae and results in Table 4 reveal that as one averages PEDRs to form higher level AC scores, the contribution of variance compo- nents involving the measurement facet being averaged across is reduced.

For example, if one averages PEDRs for a given dimension across exercises to form a dimension-level score, one would divide all variance components that have exercises as a facet by *n*e (i.e., the number of exercises across which PEDRs are averaged) to arrive at variance component estimates for dimension-level scores. For dimension-level scores, this rescal- ing effectively reduces the contribution of all components that

include exercises as a facet, relative to other components in the model. As Table 4 reveals, this rescaling can greatly change the relative contribution of components to observed variance (often of concern in investigation of the construct validity of AC scores; e.g., Kuncel & Sackett, 2012), as well as reliable and unreliable variance. For instance, note how assessee main ef- fects account for more of the reliable variance in dimension- level scores and even more of the reliable variance in overall AC scores relative to unaggregated PEDRs. This reflects the fact that one is diminishing the contribution of exercise-related components (in the case of dimension-level scores) and dimension- and exercise-related components (in the case of overall AC scores) through aggregation.

Table 4

*Decomposition of Observed Variance in Assessor Ratings as a Function of Level of Aggregation*

Postexercise dimension rating Dimension-level scores Exercise-level scores Overall assessment center score

Variance component

Formula

Between candidate variance (%)

Between- candidate subtotal

(%) Formula

Between candidate variance (%)

Between- candidate subtotal

(%) Formula

Between candidate variance (%)

Between- candidate subtotal

(%) Formula

Between candidate variance (%)

Between- candidate subtotal (%)

Reliable variance

2 2 17.2 22.4 2 33.7 45.5 2 27.9 31.1 2 51.0 57.3

candidate c c c c

2 2 1.1 1.4 2 2.1 2.9 2 /nd 0.2 0.2 2 /nd 0.4 0.5

candidate dim cd cd

cd cd

2 2 35.2 45.8 2 /ne 22.9 31.0 2 57.0 63.4 2 /ne 34.7 39.0

candidate ex ce ce

ce ce

2 2

23.4 30.4 2

/ne 15.2 20.6 2

/nd 4.7 5.3 2

/nde 2.9 3.2

candidate dim ex

cde

cde

cde

cde

Subtotal 76.8 73.9 89.8 89.1

Unreliable variance

|  |  |  |
| --- | --- | --- |
|  | 3.0 | 27.1 |
| /nd  da | 2.3 | 20.8 |
| /ne  ea | 2.1 | 19.1 |

2 2 1.0 1.3 2 2.0 7.5 2 1.6 15.9 2

candidate assessor ca ca

ca ca

2 2 6.1 8.0 2 12.0 46.0 2 /nd 1.2 12.2 2

candidate dim assessor

cda

cda

cda c

2 2 2.1 2.8 2 /ne 1.4 5.3 2 3.4 33.6 2

candidate ex assessor

cea

cea

cea c

2 2 11.8 15.3 2 /ne 7.7 29.5 2 /nd 2.4 23.4 2 /nde 1.5 13.3

4-way interaction residual

cdea

cdea

cdea

cdea

2 2 0.5 0.7 2 1.0 3.9 2 0.9 8.3 2 1.6 14.2

assessor a a a a

2 2 0.8 1.0 2 1.5 5.8 2 /nd 0.2 1.5 2 /nd 0.3 2.6

assessor dim ad ad

2 2 2

ad ad

2 2

0.2 0.3 /ne 0.2 0.6

0.4 3.9 /ne 0.2 2.2

assessor ex

2

ae

2

ae

0.6 0.7 2

ae

/n2 0.4 1.4 2

ae

/nd 0.1 1.1 2

/nde 0.1 0.6

assessor dim ex

ade

ade

ade

ade

Subtotal 23.1 26.1 10.2 10.9

PUTKA AND HOFFMAN

124

Expected reliability 0.77 0.74 0.90 0.89

*Note*. Variance components reported in this table are based on weighted mean averages of variance components across the three analysis samples. The formulae referenced above are based on the raw estimated variance components. *n*e no. of postexercise dimension ratings (PEDRs) aggregated to form a dimension-level score (i.e., no. of exercises that assess a given dimension); we assumed *n*e 3 for purposes of calculation. *n*d no. of PEDRs aggregated to form an exercise-level score (i.e., no. of dimensions assessed by a given exercise); we assumed *n*d 8 for purposes of calculation. *n*de total no. of PEDRs aggregated to form an overall assessment score; we assumed *n*de 24 for purposes of calculation. Dim dimension; ex exercise; expected reliability expected single-rater reliability.

These variance components contribute to unreliable variance because candidates are not fully crossed with assessors.

It is important to note that that the percentages of reliable and unreliable variance reported in Table 4 still assume one is defining reliable and unreliable variance from the perspective of interrater reliability. That is, they assume one is only interested in general- izing the resulting scores (regardless of level of aggregation) to new sets of assessors. As noted earlier, in cases where exercises are designed to assess similar aspects of dimensional performance (i.e., more or less viewed as parallel forms), one may not only be interested in generalizing scores across assessors but also across exercises. Given this possibility, Table 5 provides an example of

how changing the generalizations one wishes to make regarding scores changes which components actually contribute reliable and unreliable variance. Rules for determining which components con- tribute to reliable and unreliable variance as a function of desired generalizations are provided by the G-theory literature (e.g., Bren- nan, 2001, pp. 101–103).

As Table 5 illustrates, changing the generalizations one desires to make regarding AC scores impact which components actually constitute reliable and unreliable variance, and changing the level to which PEDRs are aggregated impacts how the contribution of

Table 5

*Composition of Reliable and Unreliable Variance in Assessor Ratings as a Function of Level Aggregation and Desired Generalization*

Level of aggregation/Desire to generalize scores to new . . .

Components of variance

Reliable Unreliable

Expected

reliability Interpretation

PEDR

Assessors c, cd, ce, cde ca, cda, cea, cdea, a , ad , ae , ade 0.77 Expected correlation between PEDRs for a given dimension-exercise combination provided by two different assessors

Exercises c, cd, ca, cda, a , ad ce, cde, cea, cdea, ae , ade 0.27 Expected correlation between PEDRs for a given dimension as assessed in two different exercises by the same assessor

Assessors and exercises c, cd ce, cde, ca, cda, cea, cdea, a , ad , ae , ade

Dimension

Assessors c, cd, ce/ne, cde/ne ca, cda, cea/ne, cdea/ne, a , ad , ae/

ne, , ade/ne

Exercises c, cd, ca, cda, a , ad ce/ne, cde/ne, cea/ne, cdea/ne, ae/ne, , ade/ne

Assessors and exercises c, cd ce/ne, cde/ne, ca, cda, cea/ne, cdea/ne, a , ad , ae/ne, , ade/ne

0.18 Expected correlation between PEDRs for a given dimension as assessed in two different exercises by two different assessors

0.74 Expected correlation between average PEDRs for a

given dimension (based on an average across ne exercises) provided by two different assessors

0.52 Expected correlation between average PEDRs for a

given dimension provided by one assessor (based on an average across ne exercises) and average PEDRs for that same dimension based on a new set of ne exercises

provided by that same assessor

0.36 Expected correlation between average PEDRs for a

given dimension provided by one assessor (based on an average of ratings across ne exercises) and average PEDRs for that same dimension based on a new set of ne exercises provided by a different assessor

*Note*. Expected reliabilities are based on weighted mean averages of variance components across the three analysis samples. c candidate; d dimension; e exercise; a assessor; ne no. of postexercise dimension ratings (PEDRs) aggregated to form a dimension-level score (i.e., no. of exercises that assess a given dimension); we assumed ne 3 for purposes of calculation. Note that if ne is set equal to 1, the expected reliability of the dimension-level scores equal the expected reliability of PEDRs holding the desired generalization constant.

These variance components contribute to unreliable variance because candidates are not fully crossed with assessors.

those components are scaled. Regardless of level of aggregation, components that involve Assessee Exercise interactions, but do not involve assessors, are treated as reliable variance if one wishes to generalize scores only across assessors (i.e., they are source of consistency in assessor ratings). In contrast, all components that involve Assessee Exercise interactions are treated as unreliable variance if one wishes to generalize scores across exercises or exercises and assessors (i.e., they are a source of inconsistency in ratings across exercises). The net result of these differences is that PEDRs, as well as dimension-level scores formed through aggre- gation of PEDRs, are far less reliable if one desires to generalize such scores across exercises, rather than assessors only. Thus, when estimating or describing the reliability of AC scores, it is important that researchers and practitioners clarify the generaliza-

tions they desire to make based on their scores because it can make a substantial difference in the level of reliability observed and how they calculate their estimates.

**Discussion**

In the introduction to this article, we argued that the extant literature on AC functioning suffers from various limitations that have inhibited researchers’ ability to come to an understanding of reliable and unreliable sources of variation in assessor ratings. To help put this study’s findings in perspective relative to past em- pirical summaries of AC functioning, we juxtaposed in Table 6 the variance component estimates from the current study with the most comparable estimates based on past CFA-based summaries of AC

Table 6

*Comparison of Findings of Current Study to Past Research Regarding Variance Components Underlying Postexercise*

*Dimension Ratings*

Random effects studies Reviews based on confirmatory factor analysis

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Arthur et al. | Bowler & Woehr |  | Lievens & Conway | Lance et al. | Bowler & Woehr |  | Sample | Sample | Sample |
| (2000) | (2009) |  | (2001) | (2004) | (2006) |  | 1 | 2 | 3 |

Between-candidate variance in current study

Variable

Variance components

Reliable variance

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 24.3a,f | 5.2f | — | 14i | — | 18.8 | 14.6 | 18.2 |
| 27.0b,f | 18.0f | 34g |  | 22k | 1.4 | 0.4 | 1.4 |
| 6.8c | 32.0f | 34h | 52j | 33l | 32.9 | 40.3 | 32.8 |
| — | — | — | — | — | 23.9 | 22.2 | 23.9 |

2 candidate

2

candidate dim

2

candidate ex

2

candidate dim ex

Unreliable variance

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| — 0.0e,f  date assessor | | | — | — | — | 0.0 | 3.2 | 0.0 |
| —  date dim assessor | | — | — | — | — | 9.7 | 0.0 | 8.5 |
| —  date ex assessor | | — | — | — | — | 2.3 | 0.0 | 3.5 |
| 33.8d,f | | 44.8d,f | 32m | 34m | 45m | 9.6 | 17.3 | 9.1 |
| or | 8.1 | 0.0e,f | — | — | — | 0.4 | 0.8 | 0.4 |
| — 0.0e,f —  or dim | | | | — | — | 0.5 | 0.5 | 1.1 |

2 candi

2 candi

2 candi

2

4-way interaction residual

2 assess

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2 assess |  | | | | | | | |
| 2  assessor ex | — | 0.0e,f — | | — | — | 0.1 | 0.0 | 0.5 |
| 2  assessor dim ex | — | — | — | — | — | 0.4 | 0.6 | 0.6 |
| Average MTMM correlations | | | | | | | | |
| SD–DE | .60 | .15 | — | — | .25 | .25 | .19 | .22 |
| DD–SE | .39 | .31 | — | — | .53 | .60 | .63 | .60 |
| DD–DE | .32 | .07 | — | — | .20 | .22 | .17 | .20 |

*Note*. Values represent percentages of between-candidate variance in postexercise dimension ratings accounted for by the given component. The percentages of between-candidate variance accounted for by the variance components reported in Arthur et al. (2000) and Bowler and Woehr (2009) were calculated by omitting those components that did not contribute to between-candidate variance, summing the components that remained, dividing each remaining component by the resulting sum, and multiplying by 100. Dim dimension; ex exercise. Average multitrait–multimethod (MTMM) correlations: SD–DE same dimension– different exercise correlation; DD–SE different dimension–same exercise correlation; DD–DE different dimension– different exercise correlation. Dashes indicate component not estimable by the given study.

These variance components contribute to unreliable variance when candidates are not fully crossed with assessors. a Estimate confounded with unreliable Candidate Assessor variance. b Estimate confounded with unreliable Candidate Dimension Assessor variance. c Estimate confounded with unreliable Candidate Exercise Assessor variance. d Estimate confounded with variance stemming from multiple omitted interaction ef- fects. e Estimate uninterpretable due to misspecification of assessor identification codes when fitting the random effects model. f All estimates with superscripts a– e are potentially biased as a result on nonindependent residuals stemming from omission of several key interaction terms from the random effects model. g Estimate based on correlated dimension– correlated uniqueness (CD-CU) model and confounded with multiple other sources of reliable (candidate main effect variance) and unreliable variance (e.g., Candidate Assessor variance, Candidate Dimension Assessor variance, assessor main effect variance—assuming noncrossed design). h Estimate based on CD–CU model and confounded with unreliable Candidate Exercise Assessor variance. i Estimate based on one dimension-correlated exercises (1D-CE) model and confounded with multiple sources of unreliable variance (e.g., Candidate Assessor variance, assessor main variance—assuming noncrossed design). j Estimate based on one dimension-correlated exercises (1D-CE) model and confounded with multiple other sources of reliable (Candidate Dimension variance) and unreliable variance (e.g., Candidate Assessor variance, Candidate Exercise Assessor variance, assessor main effect variance–assuming noncrossed design). k Estimate based on correlated dimensions– correlated exercises (CD–CE) model and confounded with multiple other sources of reliable (candidate main effect variance) and unreliable variance (e.g., Candidate Assessor variance, Candidate Dimension Assessor variance, assessor main effect variance–assuming noncrossed design). l Estimate based on CD–CE model and confounded with multiple other sources of reliable (candidate main effect variance) and unreliable variance (e.g., Candidate Assessor variance, Candidate Exercise Assessor variance, assessor main effect variance—assuming noncrossed design). m Estimate confounded with variance stemming from omitted interaction effects.

functioning and the two field studies that have fitted random effects models to disaggregated PEDRs. First, notice the number of blanks in the columns summarizing past research. This illus- trates the amount of information typically omitted from past field research on components of variance in assessor ratings. Second, notice from the high level of footnoting in the table (which annotates confounds present in past research) how heavily findings from past studies would need to be caveated if researchers at- tempted to use them to make inferences about components of variance. Third, unlike the current study, all of the past studies noted in Table 6 provide a decomposition of variance in PEDRs, but do not report how that composition would change for dimension-level, exercise-level, or overall-AC scores formed through aggregation of PEDRs. Given that PEDRs are not often ultimately used in practice for decision making or feedback, this is arguably an important limitation of past work that has modeled assessor ratings (see also Kuncel & Sackett, 2012). Given these limitations, we focus on a few key observations regarding the current study’s findings as a springboard to discuss implications of this work for future AC research and practice.

**It is possible to achieve a more nuanced understanding of AC functioning.** In this study, we found the residual variance component to be far smaller than that found in past studies (see Table 6). The most salient omission from past work appears to be the set of three-way interaction terms—most notably the reliable Assessee Dimension Exercise interaction. Collectively, these interaction terms accounted for between 22.8% (Sample 2) and

36.5% (Sample 3) of between-assessee variance in PEDRs the current study. These findings suggest that a more refined under- standing of how ACs function is possible than previous research would indicate. Two key examples of the refinements offered by more careful consideration of three-way interactions are insights they provide into assessees’ dimensional performance, as well as the parsing of exercise and assessor-related effects— both of which we discuss in the following sections.

**Dimensional performance is complex.** Through the lens of traditional schools of thought on AC functioning, it may be tempt- ing to take these findings as evidence that dimensions should not be central to the development and implementation of ACs (Lance,

2008). For example, Assessee Dimension effects accounted for no more than 1.4% of between-assessee variance in PEDRs. More- over, the contribution of these effects did not substantially improve when PEDRs were aggregated to the dimension-level, accounting for only 2.1% of between-assessee variance in dimension-level scores. Indeed, a far greater source of consistency among PEDRs that shared a dimension in common was assessee main effects, which accounted for no less than 14.6% of between-assessee variance in PEDRs and 33.7% of between-assessee variance in dimension-level scores. This finding is consistent with recent results by Lance et al. (2004), who found evidence for a higher level of fit for CFA models that specified a general factor (one dimension-correlated exercises [1D–CE] model) relative a model that specified multiple dimension factors (correlated dimensions- correlated exercises [CD–CE] model) to AC data (see also Kuncel

& Sackett, 2012).

Despite the small contribution of Assessee Dimension effects to between-assessee variance, it does not mean future researchers should discount the importance of dimensions to AC functioning. Indeed, doing so would ignore the fact that 15.2% of between-

assessee variance in dimension-level scores was attributable to Assessee Dimension Exercise effects, and at least 22.2% of between-assessee variance in PEDRs was attributable to such effects. More specifically, this study provides evidence that as- sessees’ performance on any given dimension exhibits a reliable (across assessors), situationally *stable* component reflecting both Assessee Dimension effects *and* assessee main effects, and a reliable situationally *variable* component reflecting both Assessee Exercise *and* Assessee Dimension Exercise effects.

The pattern of findings described appears to be consistent with theoretical frameworks that have emerged from the personality literature offering substantive explanations for differences in trait- related behavior across situations—namely, trait activation theory (TAT; e.g., Lievens et al., 2006; Tett & Guterman, 2000), and coherence theory (e.g., Cervone & Shoda, 1999; Michel & LeBreton, 2011; Mischel & Shoda, 1995). These frameworks posit that inconsistencies in trait-related behavior (dimensional perfor- mance) across situations (exercises) are not solely error but rather reflect reliable individual differences in behavior as function sub- stantive characteristics of the situations (exercises) considered (see also Gibbons & Rupp, 2009). For instance, effectively influencing others in a group setting entails using rationale to persuade others; on the other hand, in a one-on-one role play, effectively influenc- ing others may entail offering tangible rewards and feedback to influence a subordinate. Although each of these examples has influencing others as the underlying dimension, there exist natural situationally driven contingencies that lead to differences in how effective dimension-related behavior manifests itself across exer- cises.

Historically, the AC literature has framed inconsistency in di- mensional performance across exercises as evidence of an “exer- cise effect,” but this study reveals a more nuanced perspective. One part of the inconsistency in assessees’ dimensional perfor- mance appears to be function of the exercise examined and exhib- its similar effects regardless of the dimension considered (i.e., Assessee Exercise effects), whereas another part appears to be a function of the specific dimension– exercise combination exam- ined (i.e., Assessee Dimension Exercise effects). The former component is often what is emphasized in exercise- or task-centric discussions of AC functioning (Lance, 2008), yet it only provides a partial picture of inconsistency of assessees’ dimensional per- formance across exercises. The latter component is more consis- tent with interactionist perspectives on dimensional performance and trait-related behavior (e.g., Cervone & Shoda, 1999; Hoffman,

2012; Lievens et al., 2006), yet it has largely remained hidden from view in MTMM CFA-based studies of AC functioning because it has typically been confounded with residual variance.

In light of these observations, we encourage future researchers to be wary of the fact that assessees’ dimensional performance— both the situationally stable part of it and the situationally variable part of it—is multifaceted and more nuanced than most studies of AC functioning reveal. Therefore, we encourage future researchers concerned with coming to a better understanding of AC function- ing to gather data and use modeling methods that will allow them to isolate the more nuanced components of assessees’ performance revealed here and begin to explore factors that affect their magni- tude.

**Providing PEDR-level feedback may have value.** The re- sults of this study suggest that it may be important to provide assessees with feedback on their AC performance at the level of individual PEDRs, given the amount of reliable (across-assessors) performance information that would be glossed over if one limited feedback to either the exercise or dimension levels (Lievens,

2009). For example, though providing assessees with dimension- level feedback may be common practice (Spychalski et al., 1997), it would mask meaningful differences in performance specific to each exercise (Assessee Exercise effects) and dimension- exercise combination examined (Assessee Dimension Exer- cise effects) that assessors appear to rate consistently. Moreover, the literature on performance feedback has consistently suggested specificity of performance feedback is positively related to at least short- to intermediate term performance (Kluger & DeNisi, 1996). Indeed, the current study actually provides a concrete example

of how decompositions of variance can impact decisions regarding the structure of AC feedback used in practice. Several months after the completion of each AC described in this study, candidates were provided with feedback at the level of individual D–E units (based on PEDRs), with the feedback grouped by exercise (i.e., a separate PEDR profile for exercise). That is, candidates were provided with information on their performance on each dimension measured within each exercise. Decisions regarding the structure of the feedback given to candidates reflected the pattern of variance observed in assessor ratings (i.e., strong Assessee Dimension Exercise effects and Assessee Exercise effects). Though in broader AC practice, dimension-level feedback may be more com- mon, the developers of this AC adopted the perspective that presenting feedback at the dimension level just for the sake of being consistent with common practice was not consistent with the evidence at hand, and the potential benefits of providing specific, reliable PEDR-level feedback (Hoffman, 2012; Kluger & DeNisi,

1996).

Though the present findings suggest the potential value of PEDR-level feedback relative to dimension- or exercise-level feedback in the context of ACs (i.e., preventing the masking of reliable differences in performance across D–E units), we recog- nize this is an open research area. For example, though providing PEDR-level feedback would be more specific than providing feed- back at the aggregate dimension or exercise level, it would put more of an information-processing burden on assessees. Further- more, more specificity may not necessarily lead to better learning and long-term performance (Goodman & Wood, 2004). As such, we encourage future researchers to examine the effects of different levels of specificity of AC feedback on subsequent assessee out- comes (e.g., learning and development, future performance).

**Assessor biases appear to be small relative to exercise- related effects.** Assessor biases are a frequently proposed culprit for the magnitude of exercise effects (Kolk, Born, & van der Flier,

2002; Lievens, 2009; Robie, Osburn, Morris, Etchegaray, & Ad- ams, 2000; Sackett & Dreher, 1982; Woehr & Arthur, 2003). Though past studies have looked into this issue, they have done so using designs involving PEDRs provided by one assessor (e.g., Kolk et al., 2002; Robie et al., 2000)—which as noted earlier, confound reliable and unreliable variance— or have been limited to investigations in the lab (e.g., Lievens, 2001a, 2001b). In this study, we avoided these limitations and found that myriad assessor-related effects (i.e., assessor main effects and interactions

involving assessors) accounted for far less between-assessee vari- ance in PEDRs (on average, 23.0%), compared with exercise- related effects that were consistent across assessors (i.e., Assessee Exercise and Assessee Dimension Exercise), which accounted for 56.8% of between-assessee variance in PEDRs. It is important to note that this pattern emerged even though we examined ratings made by assessors based on their own individual notes and observations and without discussing assess- ees’ performance with other assessors. Indeed, assessor-related effects should become even smaller if assessors are required to discuss their performance before proving a final rating. These findings, coupled with those of Hoyt & Kerns’ (1999) more general findings on observer-related effects, suggest exercise ef- fects reported in past AC research are not likely attributable to assessor bias.

**Level of aggregation and desired generalizations matter.** The current study revealed that the relative composition of vari- ance underlying AC scores depends on whether one is considering PEDRs, or PEDRs that have been aggregated to form dimension- level, exercise-level, or overall-AC scores. The importance of considering issues of aggregation when studying the composition of variance in AC scores was also recently noted by Kuncel and Sackett (2012). However, unlike that work (which focused on issues of construct validity), in the current study, we were able to examine the effects of aggregation on a far richer set of variance components and, in turn, avoid confounds that directly impact validity-based interpretations of AC functioning. For example, as PEDRs were aggregated to form dimension-level scores in the current study, the contribution of Assessee Dimension effects to between-assessee variance increased from 1.2% to 2.1%, and the contribution of Assessee Dimension Assessor effects in- creased from 6.1% to 12.0%. In the Kuncel and Sackett (2012) work, these two interaction effects were confounded. Therefore, although they found aggregation of PEDRs to the dimension-level increased the relative contribution “dimension effects,” we found in the current study that most of this increase appears due to unreliable dimension effects (i.e., effects that are inconsistent across assessors), rather than dimension effects that are consistent across assessors. Thus, the benefits of aggregation for dimension- centric perspectives on AC functioning discussed by Kuncel and Sackett (2012) may simply be an artifact of variance that is not replicable across assessors.

Besides reinforcing the importance of aggregation issues when studying the composition of variance underlying assessor ratings, the current study also illustrated how the generaliza- tions one desires to make affect the composition of reliable and unreliable variance, and in turn the magnitude of one’s reliabil- ity estimates. As such, future AC researchers and practitioners could potentially come to very different conclusions regarding the “reliability of AC scores” and composition of observed variance in AC scores depending on these factors. Though these notions readily follow from G theory, they have arguably not been well articulated in the AC literature. Thus, it is critical that future researchers and practitioners be clear about issues of aggregation and desired generalizations when reporting reliabil- ity estimates and describing results of variance decompositions of assessor ratings.

**Limitations, Caveats, and Extensions**

Despite the positive aspects of this study, we do want to caution the reader on a few potential limitations regarding the generaliz- ability of the current study’s findings and random effects method- ology used.

**Generalizability of findings.** First, note that in the discussion of results and implications, we took caution not to imply that the pattern of reliable variance found here would always generalize to differently designed ACs. The current study is based on only a single set of samples that involved a similar incumbent population and a similar design and implementation process, which may limit its generalizability. Given strong evidence that AC design moder- ates the magnitude of dimension and exercise effects (Lievens,

1998; Woehr & Arthur, 2003), it is important that future research examine moderators of the magnitude of the more detailed vari- ance components examined here.

Though it might be tempting to dismiss these findings as spe- cific to the ACs investigated here, there are other factors that temper the limitations. First, it is noteworthy that we found the mean same dimension– different exercise (SD–DE), different dimension–same exercise (DD–SE), and different dimension– different exercise (DD–DE) correlations from the current study were quite consistent with those underlying past CFA-based sum- maries (see the last few rows of Table 6), suggesting that these findings may generalize to a wider array of ACs. Note, however, that the pattern from Arthur et al. (2000) is different, which may reflect differences in the rating processes used in in that study. Second, a comparison of the characteristics of the ACs examined in this study to characteristics of ACs examined in the literature (see Table 7) reveals a high degree of similarity (Woehr & Arthur,

2003), with a more favorable participant-to-assessor ratio (.50) and longer assessor training than is typical.

**Random effects models.** Unlike CFA-based methods, ran- dom effects models do not provide unique decompositions for each PEDR. In essence, the random effects model examined in the current study provided an expected decomposition of vari- ance for assessor ratings of any given PEDR, regardless of dimension or exercise examined. Thus, one could potentially argue that random effects models provide less detailed infor- mation than CFA models when it comes to the composition of variance for a specific PEDR. Though researchers certainly can examine random effects model by dimension (e.g., only model PEDRs designed to measure a given dimension), or by exercise (e.g., only model PEDRs for a given exercise) in an attempt to get more tailored variance component estimates, doing so re-

sults in a confounding the variance components examined here (e.g., Lievens, 2001a). For example, if one only modeled PE- DRs for a single dimension, dimension-related effects could not be uniquely estimated and would become confounded with other substantively different effects in the model (e.g., it would become impossible to differentiate Assessee Dimension ef- fects from assessee main effects).

Additionally, as noted earlier, variance components estimated via random effects models can be viewed as related to variances associ- ated with latent factors in CFA models— but unlike the latent factors in many CFA models, random effects are defined to be uncorrelated (Woehr et al., 2012). While we do not necessarily see this as a weakness of random effects models given the clean, readily interpre- table nature of the resulting variance decomposition, it is important not to directly equate the variance components arising from these models with those arising from CFA models in which dimension or exercise factors are allowed to correlate. As such, it becomes critically important that one is careful when interpreting the meaning of vari- ance components stemming from random effects models, particularly if the components are being compared with variance estimates based on other approaches to modeling AC data.

Finally, though generally favored in the modern literature on vari- ance component estimation (e.g., Searle et al., 2006), the restricted maximum likelihood variance component estimators used in this study have some potential limitations. For example, from a practical perspective, such estimators have very high memory requirements relative to ANOVA estimators (e.g., Bell, 1985; SAS Institute, 2011). Furthermore, from a theoretical perspective, there are disagreements regarding estimation of interaction effects when fixed factors are introduced in the model (e.g., Brennan, Jarjoura, & Deaton, 1980; DeShon, 2002). Last, there are also normality assumptions associated with maximum likelihood estimators. In this study, we were largely able to minimize these limitations by (a) adopting more efficient variance component estimation procedures that are better equipped for handling large amounts of data (i.e., SAS procedure PROC HPMIXED), and (b) fitting random effects models (models with random factors only)—not mixed effects models (models with ran- dom and fixed factors). Additionally, the PEDRs that we examined were largely normally distributed, and we had little theoretical reason to believe the underlying score effects would *not* be normally distrib- uted. Nevertheless, even if the distribution of score effects departed from normality, simulation work suggests maximum likelihood vari- ance component estimators can be robust to moderate departures from normality (Marcoulides, 1990).

Table 7

*Characteristics of the Current Study Compared with Assessment Center Literature Summarized in Woehr and Arthur’s (2003) Meta-Analysis*

Characteristic Current study Woehr & Arthur (2003) studies

|  |  |  |
| --- | --- | --- |
| No. of exercises No. of dimensions Participant-to-assessor ratio Rating approach | 3–4  10–12  1-to-2 (.50) Within-exercise | *M*  4.78, *SD*  1.47  *M*  10, *SD*  5.11  *M*  1.71  63% of those reporting used a within-exercise approach |
| Assessors’ occupation | Managers/supervisors | 83% of those reporting used managers/supervisors |
| Length of assessor training  Assessment center purpose | 4 days (32 hr)  Selection/promotion | *M*  3.35 days, *SD*  3.06  81% of those reporting were for selection/ promotion |

**Conclusion**

The results of this study suggest that adopting wholly dimension- centric or exercise-centric views on AC functioning is insufficient for illuminating the composition of reliable variance and unreliable vari- ance in assessor ratings. The composition of variance underlying assessor ratings is multifaceted and characterized by far more system- atic sources of variance than previous summaries of AC functioning suggest. Furthermore, this study illustrated that the conclusions one draws regarding the composition of reliable and unreliable variance will be heavily influenced by both the degree to which PEDRs are aggregated and the generalizations one desires to make regarding the resulting scores. We encourage future research to further the work presented in this study and provide a better understanding of potential moderators of the components articulated here.

**References**

American Educational Research Association, American Psychological As- sociation, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Arthur, W., Jr., & Day, E. A. (2011). Assessment centers. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology, Vol- ume 2: Selecting and developing members for the organization* (pp.

205–235). Washington DC: American Psychological Association. Arthur, W., Jr., Day, E. A., & Woehr, D. J. (2008). Mend it, don’t end it:

An alternative view of assessment center construct-related validity evi- dence. [*Industrial and Organizational Psychology,*](http://dx.doi.org/10.1111/j.1754-9434.2007.00019.x) *1,* 105–111. doi:

10.1111/j.1754-9434.2007.00019.x

Arthur, W., Jr., Woehr, D. J., & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical reexamination of the assessment center construct- related validity paradox. *Journal of Management, 26,* 813– 835.

Bell, J. F. (1985). Generalizability theory: The software problem. [*Journal*](http://dx.doi.org/10.3102/10769986010001019) *of* [*Educational and Behaviora*](http://dx.doi.org/10.3102/10769986010001019)*l Statistics, 10,* 19 –29. doi:10.3102/

10769986010001019

Borman, W. C. (2012). Dimensions, tasks, and mixed models: An analysis of three diverse perspectives on assessment centers. In D. J. R. Jackson, C. L. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 309 –320). New York, NY: Routledge.

Bost, J. E. (1995). The effects of correlated errors on generalizability and dependability [coefficients. *Applied Psychological*](http://dx.doi.org/10.1177/014662169501900206) *Measurement, 19,*

191–203. doi:10.1177/014662169501900206

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment [center ratings.](http://dx.doi.org/10.1037/0021-9010.91.5.1114) *Journal* [*of Applied*](http://dx.doi.org/10.1037/0021-9010.91.5.1114) *Psychology, 91,* 1114 –1124. doi:10.1037/0021-9010

.91.5.1114

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

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

Brennan, R. L., Jarjoura, D., & Deaton, E. L. (1980). *Some issues con- cerning the estimation and interpretation of variance components in generalizability theory* (ACT Technical Bulletin No. 36). Iowa City: IA: American College Testing Program.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin,*

*56,* 81–105. doi:10.1037/h0046016

Cervone, D., & Shoda, Y. (1999). *The coherence of personality: Social- cognitive bases of consistency, variability, and organization*. New York, NY: Guilford Press.

Connelly, B. S., Ones, D. S., Ramesh, A., & Goff, M. (2008). A pragmatic [view of assessment center exercises and dimensions. *Industrial*](http://dx.doi.org/10.1111/j.1754-9434.2007.00022.x) *and* [*Organizational Psychology,*](http://dx.doi.org/10.1111/j.1754-9434.2007.00022.x) *1,* 121–124. doi:10.1111/j.1754-9434.2007

.00022.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, NY: Wiley.

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

DeShon, R. P. (1995, May). *Restricted maximum-likelihood estimation of variance components in generalizability theory: Overcoming balanced design requirements*. Paper presented at the 10th annual conference of the Society of Industrial and Organizational Psychology, Orlando, FL. DeShon, R. P. (2002). Generalizability theory. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations* (pp. 189 –

220). San Francisco, CA: Jossey-Bass.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Gibbons, A. M., & Rupp, D. E. (2009). Dimension consistency: A new (old) perspective on the assessment center [construct validity debate. *Journal*](http://dx.doi.org/10.1177/0149206308328504) *of Management, 35,* 1154 –1180. doi:10.1177/0149206308328504

Goodman, J. S., & Wood, R. E. (2004). Feedback specificity, learning opportunities, [and learning. *Journal of Applied*](http://dx.doi.org/10.1037/0021-9010.89.5.809) *Psychology, 89,* 809 –

821. doi:10.1037/0021-9010.89.5.809

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

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

Hoffman, B. J. (2012). Exercises, dimensions, and the great battle of Lilliput: Evidence for a mixed model interpretation of AC performance. In D. J. R. Jackson, C. L. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 281–306). New York, NY: Routledge.

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

119 –151. doi:10.1111/j.1744-6570.2009.01164.x

Hoffman, B. J., & Meade, A. (2012). Alternate approaches to understanding the psychometric properties of assessment centers: An analysis of the structure and equivalence of exercise ratings. [*International Journal of*](http://dx.doi.org/10.1111/j.1468-2389.2012.00581.x) *Selection and Assessment, 20,* 82–97. doi:10.1111/j.1468-2389.2012

.00581.x

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

Howard, A. (2008). Making assessment centers work the way they are supposed [to. *Industrial and Organizational Psychology,*](http://dx.doi.org/10.1111/j.1754-9434.2007.00018.x) *1,* 98 –104. doi:10.1111/j.1754-9434.2007.00018.x

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and [what can we do about it?](http://dx.doi.org/10.1037/1082-989X.5.1.64) *Psychological Methods, 5,*

64 – 86. doi:10.1037/1082-989X.5.1.64

Hoyt, W. T., & Kerns, M. (1999). Magnitude and moderators of bias in observer [ratings: A meta-analysis. *Psychological*](http://dx.doi.org/10.1037/1082-989X.4.4.403) *Methods, 4,* 403– 424. doi:10.1037/1082-989X.4.4.403

Jackson, D. J. R. (2012). Task-based assessment centers: Theoretical perspectives. In D. J. R. Jackson, C. L. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers* (pp. 173–189). New York, NY: Routledge.

Jackson, D. J. R., Stillman, J. A., & Atkins, S. G. (2005). Rating tasks versus dimensions in assessment centers: [A psychometric comparison. *Human*](http://dx.doi.org/10.1207/s15327043hup1803_2) *Per- formance, 18,* 213–241. doi:10.1207/s15327043hup1803\_2

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption [in analysis of variance. *Psychological*](http://dx.doi.org/10.1037/0033-2909.99.3.422) *Bulle- tin, 99,* 422– 431. doi:10.1037/0033-2909.99.3.422

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis and a preliminary feedback [intervention theory. *Psychological*](http://dx.doi.org/10.1037/0033-2909.119.2.254) *Bulletin, 119,* 254 –284. doi:10.1037/0033-2909.119.2.254

Kolk, N. J., Born, M. P., & van der Flier, H. (2002). Impact of common rater variance on construct validity of assessment center [dimension](http://dx.doi.org/10.1207/S15327043HUP1504_02) [judgments. *Human Performance*](http://dx.doi.org/10.1207/S15327043HUP1504_02)*, 15,* 325–338. doi:10.1207/ S15327043HUP1504\_02

Kuncel, N. R., & Sackett, P. R. (2012). Resolving the assessment center construct validity problem. In D. R. Jackson, & B. J. Hoffman (Chairs), *Dimension, task, and mixed-model perspectives on assess- ment centers.* Symposium at the 27th annual Society for Industrial and Organizational Psychology Conference. San Diego, CA.

Lance, C. E. (2008). Why assessment centers do not work the way they are supposed [to. *Industrial and Organizational*](http://dx.doi.org/10.1111/j.1754-9434.2007.00017.x) *Psychology, 1,* 84 –97. doi:

10.1111/j.1754-9434.2007.00017.x

Lance, C. E., Foster, M. R., Nemeth, Y. M., Gentry, W. A., & Drollinger, S. (2007). Extending the nomological network of assess- ment center construct validity: Prediction of cross-situationally con- sistent and specific aspects of assessment center performance. *Human Performance, 20,* 345–362.

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. *Jour-*](http://dx.doi.org/10.1037/0021-9010.89.2.377) *nal* [*of Applied*](http://dx.doi.org/10.1037/0021-9010.89.2.377) *Psychology, 89,* 377–385. doi:10.1037/0021-9010.89.2

.377

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N., & Smith, D. E. (2000). Assessment center exercise factors represent cross- situational specificity, not method bias. *Human Performance, 13,* 323–

353.

Lance, C. E., Woehr, D. J., & Meade, A. W. (2007). Case study: A Monte Carlo investigation of assessment center construct validity [models. *Or-*](http://dx.doi.org/10.1177/1094428106289395)[*ganizational Researc*](http://dx.doi.org/10.1177/1094428106289395)*h Methods, 10,* 430 – 448. doi:10.1177/

1094428106289395

Le, H., & Putka, D. J. (2006). Reliability. In S. G. Rogelberg (Ed.),

*Encyclopedia of industrial and organizational psychology* (Vol. 2, pp.

675– 678). Thousand Oaks, CA: Sage.

LeBreton, J. M., & Senter, J. L. (2007). Answers to twenty questions about interrater reliability and [interrater agreement. *Organizational Research*](http://dx.doi.org/10.1177/1094428106296642) *Methods, 11,* 815– 852. doi:10.1177/1094428106296642

Lievens, F. (1998). Factors which improve the construct validity of assess- ment centers: A review. [*International Journal of Selection*](http://dx.doi.org/10.1111/1468-2389.00085) *and Assess- ment, 6,* 141–152. doi:10.1111/1468-2389.00085

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

Lievens, F. (2001b). Assessors and use of assessment centre dimensions: A

fresh look at [a troubling issue. *Journal*](http://dx.doi.org/10.1002/job.65) *of Organizational Behavior, 22,*

203–221. doi:10.1002/job.65

Lievens, F. (2002). Trying to understand the different pieces of the con- struct validity puzzle of assessment centers: An examination of assessor and [assessee effects. *Journal of Applied*](http://dx.doi.org/10.1037/0021-9010.87.4.675) *Psychology, 87,* 675– 686. doi:10.1037/0021-9010.87.4.675

Lievens, F. (2009). Assessment centres: A tale about dimensions, exer- cises, and dancing bears. *European Journal of Work and Organizational Psychology, 18,* 102–121. doi:10.1080/13594320802058997

Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for under- standing assessment center convergent and [discriminant validity. *Jour-*](http://dx.doi.org/10.1037/0021-9010.91.2.247) *nal* [*of Applied*](http://dx.doi.org/10.1037/0021-9010.91.2.247) *Psychology, 91,* 247–258. doi:10.1037/0021-9010.91.2

.247

Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait– [multimethod studies. *Journal of Applied*](http://dx.doi.org/10.1037/0021-9010.86.6.1202) *Psychology, 86,* 1202–1222. doi:10.1037/0021-9010.86.6.1202

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996).

*SAS system for mixed models*. Cary, NC: SAS Institute.

Lowry, P. E. (1997). The assessment center process: New directions.

*Journal of Social Behavior & Personality, 12,* 53– 62.

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

386. doi:10.2466/pr0.1990.66.2.379

Maxwell, A. E. (1968). The effect of correlated error on reliability coef- ficients. [*Educational and Psychological Measurement,*](http://dx.doi.org/10.1177/001316446802800309) *28,* 803– 811. doi:10.1177/001316446802800309

Melchers, K. G., & König, C. J. (2008). It is not yet time to dismiss dimensions in assessment [centers. *Industrial and Organizational Psy-*](http://dx.doi.org/10.1111/j.1754-9434.2007.00023.x) *chology, 1,* 125–127. doi:10.1111/j.1754-9434.2007.00023.x

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. (2008). Further evidence for the validity of assessment center dimensions: A meta- analysis of the incremental criterion-related validity of assessment center [dimension ratings. *Journal of Applied*](http://dx.doi.org/10.1037/0021-9010.93.5.1042) *Psychology, 93,* 1042–1052. doi:

10.1037/0021-9010.93.5.1042

Michel, J. S., & LeBreton, J. M. (2011). Leadership coherence: An appli- cation of personality coherence theory to the study of [leadership. *Per-*](http://dx.doi.org/10.1016/j.paid.2010.12.018) *sonality* [*and Individual*](http://dx.doi.org/10.1016/j.paid.2010.12.018) *Differences, 50,* 688 – 694. doi:10.1016/j.paid

.2010.12.018

Mischel, W., & Shoda, Y. (1995). A cognitive–affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance [in personality structure. *Psychological*](http://dx.doi.org/10.1037/0033-295X.102.2.246) *Review, 102,* 246 –

268. doi:10.1037/0033-295X.102.2.246

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

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

981. doi:10.1037/0021-9010.93.5.959

Putka, D. J., & Sackett, P. R. (2010). Reliability and validity. In J. L. Farr

& N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 9 – 49). New York, NY: Routledge.

Robie, C., Osburn, H. G., Morris, M. A., Etchegaray, J. M., & Adams, K. A. (2000). Effects of the rating process on the construct validity of assessment [center dimension evaluations. *Human*](http://dx.doi.org/10.1207/S15327043HUP1304_2) *Performance, 13,*

355–370. doi:10.1207/S15327043HUP1304\_2

Rupp, D. E., Thornton, G. C., & Gibbons, A. M. (2008). The construct validity of the assessment center method and usefulness of dimensions as focus [constructs. *Industrial and Organizational Psychology,*](http://dx.doi.org/10.1111/j.1754-9434.2007.00021.x) *1,* 116 –

120. doi:10.1111/j.1754-9434.2007.00021.x

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling [empirical findings. *Journal of Applied*](http://dx.doi.org/10.1037/0021-9010.67.4.401) *Psychology, 67,* 401– 410. doi:10.1037/0021-9010.67.4.401

SAS Institute Inc. (2011). *SAS/STAT Version 9.3 User’s Guide*. Cary, NC: SAS Institute.

Schmidt, F. L., & Hunter, J. E. (1989). Interrater reliability coefficients [cannot be computed when only one stimulus is rated. *Journal*](http://dx.doi.org/10.1037/0021-9010.74.2.368) *of Applied Psychology, 74,* 368 –370. doi:10.1037/0021-9010.74.2.368

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons [from 26 research scenarios. *Psychological*](http://dx.doi.org/10.1037/1082-989X.1.2.199) *Methods, 1,*

199 –223. doi:10.1037/1082-989X.1.2.199

Searle, S. R., Casella, G., & McCulloch, C. E. (2006). *Variance compo- nents* (2nd ed.). New York, NY: Wiley.

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

Shore, T. H., Thornton, G. C., & Shore, L. M. (1990). Construct validity of two categories of assessment [center dimension ratings. *Personnel Psy-*](http://dx.doi.org/10.1111/j.1744-6570.1990.tb02008.x) *chology, 43,* 101–114. doi:10.1111/j.1744-6570.1990.tb02008.x

Smith, P. L., & Luecht, R. M. (1992). Correlated effects in generalizability studies. [*Applied Psychological Measurement*](http://dx.doi.org/10.1177/014662169201600302)*, 16,* 229 –235. doi:

10.1177/014662169201600302

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Spychalski, A. C., Quinones, M. A., Gaugler, B. B., & Pohley, K. A. (1997). A survey of assessment center practices in [organizations in the](http://dx.doi.org/10.1111/j.1744-6570.1997.tb00901.x) United [States. *Personnel Psychology,*](http://dx.doi.org/10.1111/j.1744-6570.1997.tb00901.x) *50,* 71–90. doi:10.1111/j.1744-

6570.1997.tb00901.x

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait [activation. *Journal of Researc*](http://dx.doi.org/10.1006/jrpe.2000.2292)*h in Personality, 34,* 397– 423. doi:

10.1006/jrpe.2000.2292

Thornton, G. C., & Rupp, D. E. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and devel- opment*. New York, NY: Psychology Press.

Tsacoumis, S. (2007). Assessment centers. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 259 –292). Mahwah, NJ: Erlbaum.

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*](http://dx.doi.org/10.1037/0021-9010.90.1.108) *Psy- chology, 90,* 108 –131. doi:10.1037/0021-9010.90.1.108

Woehr, D. J., & Arthur, W., Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of [methodological factors. *Journal of*](http://dx.doi.org/10.1177/014920630302900206) *Management, 29,* 231–258. doi:

10.1177/014920630302900206

Woehr, D. J., Putka, D. J., & Bowler, M. C. (2012). An examination of G-theory methods for modeling multitrait–multimethod data: Clari- fying links to construct validity and confirmatory factor [analysis.](http://dx.doi.org/10.1177/1094428111408616) [*Organizational Researc*](http://dx.doi.org/10.1177/1094428111408616)*h Methods, 15,* 134 –161. doi:10.1177/

1094428111408616

**Appendix**

**Summary of Assessment Center Exercises and Dimensions**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table A1  *Dimension–Exercise Mapping by Sample* |  | | | | | | | | | | | | |
|  |  | Sample 1 |  |  |  |  | Sample 2 |  |  |  |  | Sample 3 |  |
| Dimension/exercise | AN | GM | IB | RP |  | AN | GM | IB | RP |  | AN | IB | RP |
| Plan, organize, prioritize | X |  | X | X |  | X |  | X | X |  | X | X | X |
| Judgment & problem solving | X |  | X | X |  | X |  | X | X |  | X | X | X |
| Decisiveness |  |  | X | X |  |  |  | X | X |  |  | X | X |
| Oral communication | X | X | X | X |  | X | X | X | X |  | X | X | X |
| Relate to others |  | X |  | X |  |  | X |  | X |  |  |  | X |
| Lead others |  | X | X | X |  |  |  | X | X |  |  |  | X |
| Organizational awareness |  |  | X |  |  |  | X | X |  |  |  |  |  |
| Knowledge of administrative procedures |  |  | X | X |  |  |  | X | X |  |  | X |  |
| Resources management |  |  | X |  |  |  |  | X |  |  |  | X | X |
| Technical Job Knowledge 1 | X |  | X |  |  | X | X | X |  |  | X | X |  |
| Technical Job Knowledge 2 | X |  | X | X |  | X |  | X | X |  | X | X | X |
| Technical Job Knowledge 3 |  |  |  | X |  |  |  |  |  |  |  |  |  |

*Note*. Sample 1 *N*  153; Sample 2 *N*  198; Sample 3 *N*  282. X indicates the given dimension was assessed by the given exercise. Descriptions of each dimension are provided in Table A3. To protect the anonymity of the participating organization, we have masked labels for those dimensions that tapped specific technical knowledge required by the target supervisory job. AN analysis exercise; GM group meeting exercise; IB in-basket exercise; RP Role play exercise.

(*Appendix continues*)

Table A2

*Exercise Descriptions*

Exercise Description

Analysis exercise

Group meeting exercise

In-basket exercise

Role play exercise

This exercise simulated a meeting in which the candidate had to brief a higher level manager (assessor) on a coworker’s work product and respond to questions from the manager. Candidates had between 75 (Samples 1 and 2) and 90 min (Sample 3) to review the exercise materials, followed by a 40-min interaction with assessors.

In Sample 1, this exercise simulated a meeting with members of the office (assessors), and the candidate was required to give a brief speech and then respond to questions. In Sample 2, this exercise simulated a meeting with community representatives (assessors) on issues related to the organization’s interaction with the community. Candidates had 20 min to review the exercise materials, followed by a 20-min interaction with assessors.

The materials for this exercise were similar to those typically found in a supervisor’s in-basket. In Sample 1, the candidate read through the materials and then, for each item, indicated (in writing) the action that he or she would take to address the item. Next, the candidate met with the assessors to explain his or her response to each item. In Samples 2 and 3, the part of the process where assessors reviewed the candidate’s written responses was removed, and assessors simply asked the candidate orally how he or she would address each item and his or her rationale for doing so during the interactive portion of the exercise. Candidates had 2 hr and 25 min to review the exercise materials, followed by a 30- (Sample 1) to 45- (Samples 2 and 3) min interaction with assessors.

This exercise simulated a meeting with a subordinate to discuss a work product produced by the subordinate. During the preparation time for this exercise, the candidate reviewed the materials and prepared for a meeting with the subordinate employee (assessor) to discuss and provide feedback on the materials. Candidates had between 40 (Samples 1 and 3) and

50 (Sample 2) min to review the exercise materials, followed by a 30-min interaction with assessors.

Table A3

*Dimension Descriptions*

Dimension Description

Plan, organize, prioritize Identifying program priorities with regard to the organization’s goals and objectives; planning, organizing, and prioritizing the activities of the work unit to effectively accomplish those goals and objectives; and delegating activities to others. Also involves managing multiple activities simultaneously to complete all within prescribed times.

Judgment & problem solving

Identifying and evaluating information; distinguishing between relevant and irrelevant information to make logical decisions; recognizing when sufficient information has been obtained to make decisions; and perceiving the impact and implications of decisions.

Decisiveness Recognizing when action is required and making decisions without undue delay.

Oral communication Listening to and comprehending verbal communication from others; preparing oral presentations; expressing facts and ideas in a clear and understandable manner; and using effective nonverbal behavior during oral communications.

Relate to others Considering and responding appropriately to the needs, feelings, and capabilities of different people in different situations; being tactful, compassionate, and sensitive; treating others with respect; and valuing cultural diversity and other individual differences in the workforce.

Lead others Assessing the skills and abilities of others to inspire, motivate, and guide them toward goal accomplishments; and applying situational leadership techniques.

Organizational awareness Understanding congressional mandates regarding organizational programs, functions of various organizational programs and directorates, and the organization’s mission and goals.

Knowledge of

administrative procedures

Having familiarity with company forms, how to fill them out and their purposes and uses, organizational reporting requirements; and knowing the appropriate chain of command for development and approval of correspondence.

Resources management Knowing appropriate expenditure and reporting procedures of organizational funds and basic budgeting principles and arithmetic.

Technical Job Knowledge 1 Each cluster of technical job knowledge reflected distinct sets of related declarative and procedural knowledge that

Technical Job Knowledge 2

Technical Job Knowledge 3

were required to effectively perform the target supervisory job.

*Note*. The descriptions offered above represent only high-level synopses of dimensions measured over the course of the three assessment center implementations. Technically, each dimension was operationally defined for assessors via the behavioral anchors reflected on each dimension– exercise unit’s rating scale (i.e., the range of behaviors that identify various levels of performance on a given dimension within an exercise).

Received September 23, 2011

Revision received June 4, 2012

Accepted October 15, 2012