A Reevaluation of Assessment Center Construct-Related Validity

Milton V. Cahoon[1], Mark C. Bowler[2] & Jennifer L. Bowler[2]

[1] RTI International, Research Triangle Park, USA

[2] Department of Psychology, East Carolina University, Greenville, USA

Correspondence: Mark C. Bowler, Department of Psychology, East Carolina University, 104 Rawl, Greenville, NC 27858, USA. Tel: 1-252-328-0013. E-mail: bowlerm@ecu.edu

Received: January 10, 2012 Accepted: March 7, 2012 Published: May 1, 2012

doi:10.5539/ijbm.v7n9p3 URL: http://dx.doi.org/10.5539/ijbm.v7n9p3

Abstract

Recent Monte Carlo research (Lance, Woehr, & Meade, 2007) has questioned the primary analytical tool used to assess the construct-related validity of assessment center post-exercise dimension ratings (PEDRs) – a confirmatory factor analysis of a multitrait-multimethod (MTMM) matrix. By utilizing a hybrid of Monte Carlo data generation and univariate generalizability theory, we examined three primary sources of variance (i.e., persons, dimensions, and exercises) and their interactions in 23 previously published assessment center MTMM matrices. Overall, the person, dimension, and person by dimension effects accounted for a combined 34.06% of variance in assessment center PEDRs (16.83%, 4.02%, and 13.21%, respectively). However, the largest single effect came from the person by exercise interaction (21.83%). Implications and suggestions for future assessment center research and design are discussed.

Keywords: assessment center, construct-related validity, generalizability theory, multitrait-multimethod

1. Introduction

Over the past three decades, assessment centers have been steadily gaining popularity with organizations worldwide, for the purposes of employee selection and development (Eurich, Krause, Cigularov, & Thornton, 2009; Joiner, 2002; Spychalski, Quinones, Gaugler, &Pohley, 1999). Constructed from the “currencies” of exercises and dimensions (Hoffman, Melcher, Blair, Kleinmann, & Ladd, 2011), assessment centers attempt to evaluate individual performance levels on a set of job-related skills. However, research on the design and function of assessment centers has brought to light some problematic results. Of particular concern is the “construct-related validity paradox” (Arthur, Woehr, &Maldegen, 2000). Specifically, assessment centersappear to exhibit both content- and criterion-related validity while simultaneously lacking construct-related validity. That is, despite utilizing high-fidelity work simulations that provide excellent predictors of job performance (Thornton & Mueller-Hanson, 2004), the primary source of variance in post-exercise dimension ratings (PEDRs) appears to be differences in the assessment center exercises rather than differences in the dimensionson which the assessment center is based (Lance, Lambert, Gewin, Lievens, & Conway, 2004; Lance et al., 2000; Petrides, Weinstein, Chou, Furnham, & Swami, 2010). However, the results of a recent Monte Carlo study (Lance, Woehr, & Meade, 2007) call into question the accuracy of the primary analytical technique that was used in the majority of the studies that have concluded that assessment centers lack construct-related validity, namely, a confirmatory factor analysis (CFA) of a multitrait-multimethod matrix (MTMM).

In light of the strong possibility of erroneous results from a CFA of an MTMM matrix, we are left with a lack of clarity regarding the current state of assessment centers. However, there is new evidence that univariategeneralizability theory (Hartley, Rao, & LaMotte, 1978; Hemmerle& Hartley, 1973) constitutes an appropriate analytical methodology for evaluating the construct-related validity of assessment center PEDRs (Bowler &Woehr, 2008). As with a CFA of a MTMM matrix, when applied to assessment center PEDRs, generalizability theory allows dimension and exercise effects to be examined. Additionally, this method facilitates the evaluation of several other relevant factors such as the person being rated (i.e., the ratee), the assessor assigning the rating (i.e., the rater), and all of the relevant interactions. Moreover, generalizability theory does not require an iterative estimation procedure that focuses on fitting theoretical models – it simply identifies the proportion of variance associated with each component.

1.1 Study Objectives

The present study sought to clarify the assessment center construct-related validity issue by utilizing a hybrid of Monte Carlo data generation and generalizability theory. Previously published MTMM matrices were used as the population parameters for generating Monte Carlo data sets, which in turn were analyzed via generalizability theory to estimate the relative contributions of the person, dimension, and exercise effects, as well as theirrespective interactions.

2. Construct-Related Validity of Assessment Centers

The current unitarian framework of validity holds that content-, construct-, and criterion-related validity are all points along the broader spectrum of construct validation (Binning & Barrett, 1989). Content-related validity is an indicator of the similarity between the subject matter of a measure and the domain that it purportedly represents, construct-related validity is an indicator of the relationship between a measure and the theoretical concept it is intended to measure, and criterion-related validity is an indicator of the relationship between a measure and a relevant behavioral indicator of performance. The process of construct validation involves all three components and seeks to determine whether a test indeed measures what it is designed to measure, the strength of the relationship between the measure and this construct, and how readily one can draw inferences from the scores the measure produces. These three validity estimates are linked to one another and one can logically assume that the establishment of two of them necessarily asserts the existence of the third (Arthur et al., 2000). Therefore, if a measure is shown to possessboth content- and criterion-related validity, then one can reasonably assume that the measure in question also possesses construct-related validity. This logic, however, has not been shown to apply to assessment center PEDRs. In fact, research has frequently demonstrated that assessment centers display both criterion- and content-related validity while failing to demonstrate construct-related validity. More specifically, assessment centers demonstrate substantial content-related validity in that they are high-fidelity simulations of managerial work (Thornton & Mueller-Hanson, 2004) that are typically based on a job analysis that generated critical work-related performance dimensions (Thornton, 1992). Moreover, whileassessment centers have evidenced strong criterion-related validities (Arthur, Day, McNelly, &Edens, 2003; Chan, 1996), the construct-related validity of assessment centersappears to be problematic.

Theoretically, the variance in assessment center PEDRs should be largely due to the individual’s performance on the particular dimensions that are being measured. That is, aratee’s PEDR on a particular dimension should be a function of the his or her behavior relating to that dimension (i.e., a person by dimension effect) rather than the exercise from which the dimension is being measured. Previous primary studies, however, have produced inconsistent results (cf.Arthur et al., 2000; Bycio, Alvares, & Hahn, 1987; Lance, Foster et al., 2004; Lance et al., 2000). In an attempt to address these inconsistencies, three extensive meta-analytic reevaluations of assessment center construct-related validity have been conducted (Bowler &Woehr, 2006; Lance, Lambert et al., 2004; Lievens& Conway, 2001;). In one way or another, all three of these reviews utilized previously published assessment center MTMM matrices. Moreover, each one utilized CFA techniques to analyze these matrices, with each study generating somewhat different results and conclusions.

2.1 Previous Reviews of Assessment Center Construct-Related Validity

2.1.1 Lievens & Conway (2001)

In their review of assessment center construct-related validity, Lievens and Conway (2001) reanalyzed 34 MTMM matrices from 24 assessment center studies. Their analyses focused on fitting six different models to each of the MTMM matrices via a CFA. These models included (1) a correlated dimension model, (2) a correlated exercise model, (3) a correlated dimension-correlated exercise model, (4) a single dimension-correlated exercise model, (5) a direct product model, and (6) a correlated uniqueness model. The correlated dimension model represents an assessment center with PEDRs that are only influenced by the dimensions being rated, whereas the correlated exercise model represents an assessment center with PEDRs that are only influenced by the exercises utilized to evaluate the dimensions. The correlated dimension-correlated exercise model is a mixedmodel in whichassessment center PEDRs are a function of both the dimension being rated and the exercise in which they are being rated. The single dimension-correlated exercise model represents an assessment center with PEDRs that are influenced by a single dimension factor (e.g., *g*) as well as the exercises. Both the direct product and the correlated uniqueness models are statistical variations of the correlated dimension-correlated exercise model. In the direct product model the correlations between manifest variables (e.g., PEDRs) are modeled as a multiplicative function between dimensions and exercises (rather than as an additive function), and in the correlated uniqueness model the exercise effects are not explicitly modeled in favor of estimating them from the correlations among the uniquenesses.

Results from the Lievens and Conway (2001) analyses indicated that both the correlated dimension and the correlated exercise models fit the data poorly, providing an adequate fit for only 3% and 29% of the matrices, respectively. The single dimension-correlated exercise model performed somewhat more favorably, producing an acceptable fit for 53% of the matrices. In contrast, both the correlated dimension-correlated exercise model and direct product model demonstrated acceptable fit with 85% and 81% of the data, respectively. However, the correlated uniqueness model emerged as the best performance model, fitting 88% of the MTMM matrices. Based on the conclusion that the correlated uniqueness model provided the most appropriate fit for the data, Lievens and Conway estimated that the mean proportion of variance that was attributable to dimensions was .34 and that the mean proportion of variance that was attributable to exercises was also .34. Furthermore, they noted that these values varied greatly and that several models featured highly intercorrelateddimensions, with the average dimension intercorrelation being .71. Nonetheless, they concluded that dimensions play a greater role in assessment center ratings thanwas previously suspected, but the study fell short of demonstrating that dimensions have a greater impact than the exercises from which they are drawn.

2.1.2 Lance, Lambert et al. (2004)

Due to several problematic statistical issues with the LievensandConway (2001) review, particularly concerns with the utilization of the correlated uniqueness model, Lance, Lambert et al. (2004) reexamined theLievens and Conway (2001) data as well as five additional MTMM matrices (for a total of 39 MTMM matrices). As with the Lievens and Conway review, Lance, Lambert et al. individually analyzed each of the MTMM matrices via a CFA. However, in their evaluation, only three models were tested: (1) the correlated dimension-correlated exercise model, (2) the correlated exercise model, and (3) the single dimension-correlated exercise model. Overall, their results indicated that the correlated dimension-correlated exercise model provided a good fit for only two of the 39 MTMM matrices (5%) and the correlated-exercise model provided a good fit for only 2% of the matrices. In contrast, the single dimension-correlated exercise model produced admissible solutions for a substantial number of the MTMM matrices (49%) with a mean dimension factor loading of .14 and a mean exercise factor loading of .52. Thus, they concluded that exercise effects preside over dimension effects in assessment center PEDRs.

2.1.3 Bowler and Woehr (2006)

Bowler and Woehr (2006) conducted a similar study that also reexamined a similar set of previously reported assessment center MTMM matrices. However, rather than individually reanalyze each MTMM matrix, they chose to recode the data from 35 MTMM matrices into a single MTMM matrix comprised of six dimensions and six exercises. They then conducted a CFA on this single MTMM matrix to assess the fit of several different analytical models: (1) thecorrelated dimension model, (2) thecorrelated exercise model, (3) thesingle dimension-correlated exercise model, (4) asingle dimension-uncorrelated exercisesmodel, (5) thecorrelated dimension-correlated exercise model, and (6) thecorrelated dimension-uncorrelated exercise model. Their results noted that, with the exception of the correlated dimension model, all of the models demonstrated a reasonable fit; however, the correlated dimension-correlated exercise model was marginally superior, generating a mean dimension factor loading of .47 and a mean exercise factor loading of .58. Thus, contrary to the findings of Lance, Lambert et al. (2004), Bowler and Woehr concluded that exercise effects do not necessarily take precedence over dimension effects.

Overall, these three aforementioned reviews produced notably different results. Whereas results from the Lance, Lambert et al. (2004) analysis sided with other individual studies that concluded that assessment center PEDRs are a function of exercises and not dimensions (cf.Bycio et al., 1987; Lance et al., 2000; Lance, Foster, et al., 2004; Lance et al., 2007; Lievens& Conway, 2001), Bowler and Woehr (2006) and Lievens and Conway (2001) produced results to contrary. Interestingly, the findings of all of these studies were based on the same statistical technique – a CFA of a MTMM matrix. Unfortunately, the suitability of this method has recently been called into question.

2.2 Confirmatory Factor Analysis and MTMM Matrices

To assess the appropriateness of conducting a CFA of an MTMM matrix, Lance et al. (2007) conducted a Monte Carlo evaluation of this procedure. For this study, they generated three population models,each representing a different model of assessment center functioning found in previous assessment center research. Each of these models was similar in that eachmodeled an assessment center comprised of five dimensions, each of which was measured in three exercises. However, the models differed with regard to the nature of their latent structure. The first two models represented the correlated dimension-correlated exercise(CDCE) and single dimension-correlated exercise (1DCE) models. The third model was based on an uncorrelated dimension-correlated exercise model that also took into account an overall person effect (UDCE+*g*). Using these three models as population parameters, Lance et al. (2007) generated 500 sample MTMM matrices for each model. A CFA was then conducted to assess the fit of each of the three population models (i.e., CDCE, 1DCE, UDCE+*g*) to each of the sample MTMM matrices. For each model, they noted whether the model converged within 1,000 iterations and whether it produced an admissible solution.

When fitting the three population models (i.e., CDCE, 1DCE, UDCE+*g*) to the sample data, Lance et al. (2007) noted several problematic results. First and foremost, when the CDCE model converged to an admissible solution, it fit only 61% of the CDCE data. Thus, for 39% of the CDCE data, the CFA was unable to generate an appropriate solution. However, when the 1DCE model was applied to the CDCE data it converged to an admissible solution for 100% of the data. Similarly, when fitting the population models to the UDCE+*g* data, the UDCE+g model produced an admissible solution for only 52% of the data. However, when the 1DCE model was applied to the UDCE+*g* data, it converged to an admissible solution for 99% of the data. In contrast, when the 1DCE sample data was examined, the CDCE model converged to an admissible solution for only 1% of the data and the UDCE+*g* model converged to an admissible solution for only 10% of the data. However, the 1DCE model converged to an admissible solution for 100% of the data.

Overall, the findings of Lance et al. (2007) suggest that the results produced by a CFA of an MTMM matrix may be inaccurately biased towards the 1DCE model. That is, regardless of the true nature of the data, the 1DCE model is most likely to emerge as the appropriate model. Thus, the conclusions of the past 20 years of research regarding the construct-related validity of assessment centers, and more specifically, the results of numerous studies that have concluded that assessment center PEDRs are best represented by the 1DCE model (e.g., Bycio et al., 1987; Fleenor, 1996; Lance, Foster, et al., 2004; Lance, Lambert, et al., 2004; Lance et al., 2000; Schneider & Schmidt, 1992), may include some erroneous results.

2.3 Applying Generalizability Theory to Assessment Center Ratings

A solution for avoiding the problems associated with a CFA of an MTMM matrix is the application of generalizability theory to the evaluation of assessment center PEDRs. Generalizability theory (i.e., variance partitioning) examines the different sources of variance associated with PEDRs and estimates the relative impact that each source has on the ratings (Cronbach, Gleser, Nanda, &Rajaratnam, 1972). Bowler and Woehr (2008) have asserted that this method is better suited for examining the construct-related validity of assessment centers in that (1) it takes into account sources of variance that cannot be assessed with the traditional CFA of an MTMM matrix, and (2) it generates results that are representative of the population data.

2.3.1 Novel Sources of Variance

Three primary sources of variance may be readily assessedvia the application of generalizability theory toassessment center PEDRs: the person being rated, the dimension being rated, and the exercise from which the rating was made. The dimension effect represents the variance in PEDRs that can be attributed to certain dimensions receiving overall higher or lower ratings in comparison with other dimensions, whereas the exercise effect represents the degree to which certain exercises engender higher/lower PEDRs across all dimensions. The person effect represents a general performance factor that is not due to interactions with the dimension effect or exercise effect.

In addition to the three main effects, the generalizability theory approach evaluatesthree relevant interaction effects. The person by dimension interaction effect represents the amount of variance attributed to individuals scoring higher/lower on certain dimensions regardless of the exercise being employed. This is indicative of cross-situational consistency, which represents the degree to which a person’s dimension ratings are consistent across exercises and serves as an indicator of construct-related validity. In contrast, the person by exercise interaction effect represents the amount of variance attributed to a person receiving generally high/low scores on certain exercises, regardless of the dimension being measured. This provides evidence of situational specificity, meaning that it represents the degree to which a person does not score consistently on dimension ratings across exercises. Situationalspecificity is counterintuitive to assessment center functioning because it suggests that a person’s dimension ratings are largely a function of their performance on a particular exercise rather than the dimension that was intended to be measured (Lance et al., 2000). Lastly, the dimension by exercise interaction effect represents the amount of variance attributed to a specific dimension being measured in a specific exercise. This effect has previously been examined by Lievens et al. (2006) and concerns the observability of a particular dimension in a particular exercise.

Overall, examining these sources of variance provides substantially more information regarding the functioning of assessment centerPEDRs than does traditional CFA. Traditional CFA methods only provide a rudimentary view based on dimension and exercise effects. These two effects – along with their noted levels of intercorrelation – do not provide much detail regarding what is occurring with the PEDRs. Dimension effects are regarded as "good" and exercise effects are "bad", but little additional information is provided. In contrast, the generalizability theory approach provides more detailed information regarding the dimension and exercise effects. For example, what would simply be described as a dimension effect via CFA can be decomposed into (1) a pure dimension effect that comes from different dimensions being systematically rated differently across each other, (2) a person by dimension interaction that stems from differences in individuals' performance on the dimensions, and (3) a dimension by exercise interaction that comes from differences in ratings based on the exercise from which it is drawn.

2.3.2 Monte Carlo Support

In their Monte Carlo examination of assessment center PEDRs, Bowler and Woehr (2008) generated four population models. The first was an *optimal* model that featured low exercise loadings and high dimension loadings. The second was a *mixed* model with both high dimension and high exercise loadings. The third was a *worst-case* model that had high exercise loadings and low dimension loadings. The fourth and final population represented a *null* model that featured low dimension loadings and low exercise loadings. For each of these models, Bowler and Woehr generated 500 sample data sets. This data was then subjected to generalizability theory analyses. Overall, the analyses from Bowler and Woehryielded appropriate results for each population model that was analyzed. For example, the primary source of variance in the *optimal* model was the person by dimension effect with 29.79%. In comparison, for the *worst-case* model, the primary source of variance was the person by exercise effect with 27.35%. Thus, they concluded that the application of univariate generalizability theory is an accurate and appropriate method for evaluating assessment center functioning.

2.3.3 Previous Empirical Applications

Generalizability theory has rarely been applied to examinations of assessment center construct-related validity. Arthur et al. (2000) first applied this technique to assessment centers when they examined four sources of variance in assessment center PEDRs as well as the relevant two-way interactions. Specifically, they evaluated the variance attributed to the person, dimension, exercise, rater, and person by dimension, person by exercise, and dimension by exercise interaction effects. Their results indicated that the person, dimension, and person by dimension effects accounted for a substantial amount of the total systematic variance in assessment center PEDRs (59%). Specifically, the dimension main effect accounted for roughly 21% of the total variance, the person effect accounted for 18%, and the person by dimension effect accounted for 20% of the total variance. In contrast, the exercise effect accounted for less than 1% of the variance, and the person by exercise effect accounted for roughly 5% percent of the total variance.Similarly, Jackson et al. (2005) examined the relative variance accounted for by person, dimension, and exercise factors (including the associated two-way interactions). Their results indicated that that the person effect, dimension effect, and person by dimension interaction accounted for approximately 36% of the variance (31.9%, 2.2%, and 1.8%, respectively). However, Jackson et al. also noted that approximately 37% of the variance in assessment center PEDRs was attributable to a combination of the exercise effect (3.2%) and the person by exercise interaction (34%). Most recently, Bowler &Woehr (2009) examined the variance accounted for by person, dimension, exercise, and rater factors. They noted a substantial person by exercise effect (28.4%) as well as a strong person by dimension effect (16.0%), and moderate dimension and person effects (6.7% and 4.6%, respectively).

All of these studies are relatively unique in that they provide considerable support for the expected person, dimension, and person by dimensions effects (i.e., those effects associated with the object of measurement and considered supportive of construct-related validity). However, results of these three studies diverge with respect to variance estimates associated with exercise and person by exercise effects (i.e., those effects traditionally considered unsupportive of assessment center construct-related validity). Unfortunately, these are the only three studies to apply generalizability theory to assessment center PEDRs from functioning assessment centers (i.e., not experimental data). Thus, there is a relatively small sample of studies that do not rely solely on a CFA of an MTMM matrix from which to base any conclusions regarding the overall nature of the construct-related validity of assessment centers.

3. Reevaluating Assessment Center Construct-Related Validity

Despite the plethora of studies conducted on the assessment center construct validity paradox, few have chosen to utilize generalizability theory to examine what role exercises and dimensions play in assessment center

PEDRs (Arthur et al., 2000; Jackson, Stillman, & Atkins, 2005). The present study seeks to rectify this shortcoming by applying a hybrid of Monte Carlo data generation and generalizability theoryin an effort to summarize the current state of assessment center without utilizing a problematic methodology. Specifically, we sought to apply generalizability theory to the substantial number of assessment centers that have been previously studied (e.g., Arthur et al., 2001; Bycio et al., 1987; Schneider & Schmitt, 1992). However, in most cases it is impossible to retrieve the necessary raw data from these studies – only the published MTMM matrices are available. Thus, for each MTMM matrix that met our inclusion criteria, we generated 500 sample datasets based on the MTMM matrix. Each of these datasets was then subjected to a generalizability theory analysis, and the results were aggregated to provide general information about the likely nature of that particular assessment center.

3.1 Assessment Center Design Features

In addition to the primary analyses, analyses were also conducted to evaluate the relationship between particular assessment center design features and differences in sources of variance. In particular, we examined differences in variance components in relation to (1) the total number of dimensions assessed, (2) a crossed versus non-crossed design, (3) the use of behavioral checklists, (4) the purpose of the assessment center (selection versus development), and (5) the occupation of the assessor.

3.1.1 Number of Dimensions

The number of dimensions rated by an assessment center has previously been shown to be vital to the construct-related validity of assessment center PEDRs (Woehr& Arthur, 2003). Traditionally, it is held that the greater the number of dimensions that are utilized by an assessment center, the more difficult it becomes for a rater to distinguish between said dimensions (Gaugler& Thornton, 1989). Thus, assessment centers with five or fewer dimensions should demonstrate superior construct-related validity.

> *Hypothesis 1: Assessment centers with five or fewer dimensions will have significantly higher person, dimension, and person by dimension effects,and significantly lower exercise and person by exercise effects, than assessment centers with more than five dimensions.*

3.1.2 Crossed vs. Non-crossed Design

Assessment centers that utilize a fully crossed design are those in which every dimension included in the assessment center is measured in each exercise(Thornton & Mueller-Hanson, 2004). In contrast, assessment centers with a non-crossed design typically include only a particular subset of dimensions in each of the exercises. For example, an assessment center might feature an in-basket exercise that examines three dimensions (analysis, judgment and delegation), a role-play exercise that only measures two dimensions (analysis and confrontation), and a leaderless group discussion exercise that measures three dimensions (analysis, judgment, and confrontation). The fundamental nature of assessment center design – the measurement of multiple traits via multiple methods – is aligned with the traditional construct-related validity model of Campbell and Fiske (1959). Thus, the fully crossed design should demonstrate superior construct-related validity.

> *Hypothesis 2: Assessment centers utilizing a fully crossed design will have significantly higher person, dimension, and person by dimension effects, and significantly lower exercise and person by exercise effects, than assessment centers with more than five dimensions.*

3.1.3 Behavioral Checklists

A behavioral checklist is a tool utilized in some assessment centers to assist raters in assigning their ratings. Behavioral checklists allow raters to simply indicate if any of a set of particular behaviors was exhibited by the assessee during the exercise. Typically, behavioral checklists are employed in an attempt to reduce the cognitive load on assessors and to improve the accuracy of their ratings (Donahue, Truxillo, Cornwell, &Gerrity, 1997). Thus, the use of behavioral checklists should improve the construct-related validity of assessment center ratings.

> *Hypothesis 3: Assessment centers that utilize behavioral checklists will have significantly higher person, dimension, and person by dimension effects, and significantly lower exercise and person by exercise effects, than assessment centers with more than five dimensions.*

3.1.4 Assessment Center Purpose

Selection-oriented assessment centers are used to distinguish individuals who are best suited for a particular position based on their overall assessment center performance, whereas developmental assessment centers are used to examine the strengths and weaknesses of current employees in an attempt to highlight their weaknesses and facilitate their improvement. Traditional thinking holds that developmental assessment centers should have clearer dimensions, in that their purpose is to provide dimension-level feedback (Kudish, Ladd, & Dobbins, 1997; Woehr& Feldman, 1993). Thus, developmental assessment centers should demonstrate superior construct-related validity than those that are designed for selection purposes.

> *Hypothesis 4: Assessment centers that are intended for developmental purposes will have significantly higher person, dimension, and person by dimension effects, and significantlylower exercise and person by exercise effects than assessment centers that are designed for selection purposes.*

3.1.5 Assessor Occupation

Traditionally, assessment center employees are classified as either professional assessors or managers employed by the organization utilizing the assessment center. Managers are raters that typically come from the within the organization in which the assessment center is being used. They may be supervisors or individuals who are knowledgeable in terms of the job being assessed (e.g., subject matter experts). In contrast, professional assessors generally come from outside the organization and have extensive knowledge of assessment center functioning and the behaviors that are representative of the dimensions being assessed. Previous research has suggested that ratings made by professional assessors are typically superior to those of managers (Woehr&Arthus, 2003).

> *Hypothesis 5: Assessment centers that employ professional assessors will have significantly higher person, dimension, and person by dimension effects, and significantly lower exercise and person by exercise effects, than assessment centers that employ managers employed by the organizations assessors.*

4. Method

4.1 Literature Search and Inclusion Criteria

To retrieve the data necessary for the analyses, a search of the relevant online databases was conducted (e.g., PsycINFO, PsycARTICLES, Dissertation Abstracts, etc.). Search terms included *assessment center* in conjunction with *multitrait-multimethod* and *construct validity*. In addition to searching the relevant databases, studies that were included in the previous reviews (i.e., Bowler &Woehr, 2006; Lance, Lambert, et al., 2004; Lievens& Conway, 2001) were examined for potential sources of data.

In order to be included in the analyses, studies needed to provide an assessment center based MTMM matrix as well as the means and standard deviations of each variable. Unlike previous assessment center meta-analyses (e.g., Lance et al., 2004) that utilized MTMM matrices, means and standard deviations were necessary to estimate the dimension and exercise effects so that we could provide the most accurate analysis of each assessment center. Subsequently, some more recently published MTMM matrices were not included in our analyses (e.g., Lance, Foster, Nemeth, Gentry, &Dollinger, 2007). Additionally, the studies' MTMM matrices had to (1) come from an assessment centers in which ratings of individual dimensions were collected for each exercise (i.e., PEDRs) and (2) report the sample size on which the correlations were based. When any of this information was missing (e.g., if a study provided only mean correlations across dimensions and/or exercises), the authors were contacted in an attempt to obtain the full matrix. Overall, 14studies dating from 1987 to 2004 and reporting 23 unique MTMM matrices met the inclusion criteria. Of the 23 MTMM matrices, 52% were included in the Lievens and Conway (2001) review, 61% were included in the Lance, Lambert, et al. (2004) review, and 65% were included in the Bowler and Woehr (2006) review. Table 1 presents the complete list of studies providing matrices for this study as well as several of the assessment centers' design features.

Table 1. Study demographics

Study	Sample size	No. of dimensions	No. of exercises	Fully crossed design?	Behavioral checklists?	AC purpose	Assessor occupation
Arthur et al. (2000)	149	4	3	Yes	No	Developmental	Managers
Becker (1990)							
1	48	5	4	No	No	Selection	Managers
2	48	5	4	No	No	Selection	Managers
3	48	5	4	No	No	Selection	Managers
4	48	5	4	No	No	Selection	Managers

Bycio et al. (1987)	1170	8	5	Yes	No	Both	Managers
Chorvat (1994)	207	11	4	No	Yes	Developmental	Assessors
Fredericks (1989)	66	8	3	No	No	Selection	Managers
Jansen & Stoop (2001)	581	4	2	Yes	No	Selection	Managers
Kolk et al. (2002)							
1	100	3	2	Yes	No	Selection	?
2	100	3	2	Yes	No	Selection	?
Kolk et al. (2003)							
1	149	3	4	Yes	No	Developmental	Assessors
2	149	3	4	Yes	No	Developmental	Assessors
3	690	3	2	Yes	No	Developmental	Assessors
4	690	3	2	Yes	No	Developmental	Assessors
Kudisch et al. (1997)	138	6	4	No	No	Developmental	Assessors
Lance et al. (2004)							
1	291	6	3	No	Yes	Promotional	Managers
1	206	6	3	No	Yes	Promotional	Managers
Lievens et al. (2003)	166	7	3	No	Yes	Selection	Assessors
Parker (1991)	379	11	3	No	No	Developmental	Managers
Robbie et al. (2000)							
1	100	4	2	Yes	Yes	Developmental	Assessors
2	100	4	2	Yes	Yes	Developmental	Assessors
Schneider & Schmitt (1992)	89	6	2	Yes	Yes	Developmental	Assessors

Notes. Several other factors are often included in AC reviews (e.g., participant-to-assessor ratio, transpartent dimensions, exercise similarity, assessor training, length of training). However, do to a lack of information regarding these facets they could not be included in the moderator analyses and are thus not included in this table. ftc = MTMM matrices that failed to converge to an admissable solution in the Lance et al. (2004) analyses. *MTMM matrices not included in the Lance et al. (2004) analyses.

4.2 Data Generation

Similar to the Lance et al. (2007) Monte Carlo study, we treated MTMM matrices as population matrices that served as the basis from which we generated data. However, unlike Lance et al., we utilized MTMM matrices that were drawn from actual assessment centers. Furthermore, rather than generate additional MTMM matrices for CFA-based analyses, we used each of the MTMM matrices as the basis for generating 500 sample data sets. To generate the data, a Cholesky decomposition was first calculated for each MTMM matrix. A set of random, normally distributed numbers were then multiplied by the Cholesky matrix. Each set contained the same number of variables as the original MTMM matrix with the number of sets being equal to the original sample size of the particular assessment center. Thus, the sample data sets were each created with the sample size reported in the original study and were comprised of the appropriate dimensions and exercises. For example, each of the 500 data sets that were generated based on the Bycio et al. (1987) MTMM matrix had a sample size of 1170 and 40 variables representing the combination of eight dimensions and five exercises in their fully-crossed design.

4.3 Data Analyses

Each sample data set was analyzed via the SAS VARCOMP procedure with the MIVQUE0 method (Hartleyet al., 1978).The MIVQUE0 method was chosen over traditional Maximum Likelihood or Type 1 methods due to the substantial difference in computational time. The MIVQUE0 method is one of the most efficient computational methods available (Bell, 1985; Brennan, 2001). For one sample dataset, the average MIVQUE0 analysis runtime was less than one second. In comparison, the average maximum likelihood analysis runtime for one sample dataset was approximately one minute. Thus, given the substantial number of analyses that were conducted for this study (~12,000), the maximum likelihood method was not feasible. Furthermore, the MIVQUE0 method makes no assumptions regarding the normality of the data and can be utilized for analyzing unbalanced designs (i.e., assessment centers that do not fully cross dimensions and exercises; Hartleyet al., 1978). Thus, for this study we felt that this was the superior analytical technique.

All analyses were based on a three-facet (person, dimension, and exercise) random effects design with one repeated measure (assessment center ratings). In addition to the main effects, all of the two-way interactions were included (i.e., person by dimension, person by exercise, and dimension by exercise). However, not all studies assessed all dimensions in all exercises. Therefore, not all MTMM matrices that were assessed were fully crossed. However, all dimensions were assessed in at least two exercises, so dimensions were not nested within exercises. This represents a type of fractional factorial design that has been shown to be fairly robust with respect to lower-order effects in the model (Cochran & Cox, 1957; Connor & Young, 1961; Federer, 1955).

5. Results

The results for the initial simulations based on the 23 MTMM matrices are displayed in Table 2. Variance estimates were obtained which corresponded to the seven modeled effects: person, dimension, exercise, person by dimension, person by exercise, dimension by exercise, and error (i.e., variance attributable to all components not estimated in the model as well as measurement error). Overall, an initial examination of the variance estimates indicates a fairly complex pattern of results. Specifically, those components associated with the person and dimensionmain effects accounted for a substantial proportion of the total variance (34.06%). Overall, this variance was distributed across the person main effect (16.83%), the dimension main effect (4.02%), and the person by dimension interaction (13.21%). Moreover, the variance associated with the exercise main effect was minimal (2.60%), as was the variance attributed to the dimension by exercise interaction (4.17%). However, other than error, the largest effect was accounted for by the person by exercise interaction (21.83%).

Overall, this suggests that, across assessment centers, a significant proportion of the variance in assessment center ratings is in fact associated with the object of measurement (person, dimension, and person by dimension), with a substantial proportion of this variance attributed to the person by dimension interaction. That is, aconsiderable amount of variance is associated with facets that are supportive of construct-related validity and proper assessment center functioning. However, a substantial proportion of the variance in ratings is also attributable to a person by exercise interaction. Thus, along with the "good" variance, there is also a considerable amount of "bad" variance in assessment center PEDRs – or, at the very least, variance that supports the situational specificity issue discussed by Lance et al. (2000).

Although some assessment centers appear to have PEDRs that are appropriately based on the person by dimension interaction (e.g., Fredericks, 1989; Robie, Osburn, Morris, Etchegaray, & Adams, 2000, matrix 2) others do not (e.g., Bycio et al., 1987, Parker, 1991). Similarly, although some assessment centers appear to engender a relatively small amount of situational specificity (e.g., Becker, 1990, matrix 1; Chorvat, 1994; Fredericks, 1989; Robie et al., 2000, matrix 1), others appear to engender massive amounts of situational specificity (e.g., Bycio et al., 1987; Jansen & Stoop, 2001). This finding is disturbing because it implies that true assessment center functioning varies greatly across differing assessment centers.

Further examination of these results highlights an interesting issue regarding the range of observed variance associated with each of the effects in the model. The standard deviations of the average observed variance for both the person by dimension and person by exercise interaction effects were greater than 10% (*SD* = 11.70% and 16.89%, respectively). In contrast, the standard deviations of the person, dimension, exercise, and dimension by exercise effects were lower than 10% (*SD* = 8.24%, 5.51%, 5.06%, and 3.95%, respectively). Thus, there appears to be greater similarity across assessment centers regarding the sources of variance associated with the person, dimension, exercise, and dimension by exercise effects. In contrast, the person by dimension and person by exercise effects seem to vary substantially across assessment centers.

An additional issue of note is that the Bycio et al. (1987) study produced results that differed substantially from those of the other studies examined. In particular, the Bycio et al. MTMM matrix produced a person by exercise interaction that accounted for roughly 66% of the total variance. This is considerably greater than the next highest person by exercise effect of 47% that was generated by the Jansen et al. (2001) MTMM matrix. To evaluate the impact of this potential outlier, separate means and standard deviations were calculated with the Bycio et al. study excluded. Overall, this analysis yielded only subtle differences. The components associated with the person and dimension effects accounted for a substantial proportion of the total variance (34.06%), with this variance distributed across the person main effect (17.31%), the dimension main effect (4.15%) and the person by dimension interaction (13.80%). The variance associated with the exercise main effect was minimal (2.66%), as was the variance attributed to the dimension by exercise interaction (4.35%). However, other than error, the largest effect was still attributable to the person by exercise interaction (19.84%).

Table 2. Monte Carlo simulation results

	Proportions of Variance						
Study	s_p	s_d	s_x	s_{pd}	s_{px}	s_{dx}	$s_{pdx,e}$
Arthur et al. (2000)	**23.70%**	**23.00%**	0.00%	**19.20%**	5.10%	3.70%	25.30%
Becker (1990)							
1	**11.60%**	1.50%	0.80%	**24.60%**	0.10%	2.00%	58.80%
2	**12.10%**	0.70%	0.60%	**20.80%**	2.40%	4.80%	58.60%
3	**19.00%**	0.50%	1.00%	**17.60%**	**13.90%**	1.20%	46.80%
4	**23.70%**	0.40%	0.90%	9.70%	**15.90%**	1.20%	48.30%
Bycio et al. (1987)	6.30%	1.20%	1.20%	0.30%	**65.70%**	0.20%	25.10%
Chorvat (1994)	4.10%	9.0%	0.70%	**16.40%**	3.00%	8.00%	58.80%
Fredericks (1989)	4.00%	0.40%	2.70%	**43.90%**	2.30%	6.00%	40.40%
Jansen et al. (2001)	**20.40%**	0.10%	0.50%	5.40%	**47.20%**	1.90%	24.50%
Kolk et al. (2002)							
1	**17.50%**	7.00%	0.20%	7.80%	**16.20%**	0.40%	50.90%
2	**20.10%**	2.00%	0.40%	**10.30%**	**30.20%**	1.20%	35.50%
Kolk et al. (2003)							
1	**17.30%**	2.80%	0.30%	9.60%	**24.70%**	1.20%	44.10%
2	**12.00%**	1.90%	0.40%	**11.00%**	**31.20%**	0.80%	42.80%
3	**27.80%**	0.00%	0.50%	**14.50%**	**41.30%**	3.90%	11.90%
4	**31.60%**	0.20%	0.00%	**13.40%**	**39.50%**	6.20%	8.40%
Kudisch et al. (1997)	8.30%	1.70%	1.10%	7.30%	**22.70%**	**11.30%**	47.60%
Lance et al. (2004)							
1	**19.30%**	6.60%	18.60%	2.30%	**32.40%**	5.90%	14.60%
2	**18.20%**	8.90%	**15.40%**	2.00%	**34.40%**	12.00%	8.95%
Lievens et al. (2003)	**11.00%**	**13.40%**	**10.70%**	**18.50%**	9.90%	6.80%	29.20%
Parker (1991)	4.50%	0.00%	1.20%	0.01%	**15.00%**	13.40%	65.20%
Robbie et al. (2000)							
1	**30.10%**	6.40%	2.10%	5.70%	**22.10%**	2.40%	31.00%
2	**25.30%**	3.80%	0.20%	**42.20%**	2.10%	0.40%	26.00%
Schneider & Schmitt (1992)	**19.30%**	1.00%	0.20%	1.30%	**24.90%**	1.10%	52.30%
Including Bycio et al. (1987)							
M	**16.83%**	4.02%	2.60%	**13.21%**	**21.83%**	4.17%	37.18%
SD	8.24%	5.51%	5.06%	11.70%	16.89%	3.95%	17.14%
Excluding Bycio et al. (1987)							
M	**17.31%**	4.15%	2.66%	**13.80%**	**19.84%**	4.35%	37.73%
SD	8.10%	5.61%	5.17%	11.63%	14.26%	3.94%	17.33%

Notes. p = person; d = dimension; x = exercise; pd = person by dimension; px = person by exercise; dx = dimension by exercise; pdx,e = error. Variance components greater than 10% are in boldface. The largest variance component for each MTMM matrix is underlined.

5.1 Assessment Center Design Features

As previously noted, we examined differences in variance components in relation to (1) the total number of dimensions assessed, (2) a crossed versus non-crossed design, (3) the use of behavioral checklists, (4) the purpose of the assessment center(selection versus development), and (5) the occupation of the assessor. Results from each of the 23 data sets were separated according the characteristic being examined, and t-tests were conducted to evaluate the significance of the differences (see Table 3).However, due to the massive amount of data, all differences were significant. Thus, we calculated Cohen's *d* effect sizes (Cohen, 1988) for each comparison to better judge the differences between design features. To this end, we considered effects sizes that were greater than .40 to be substantial.

Table 3. Moderators of variance partitioning results

Moderator	s_p	s_d	s_x	s_{pd}	s_{px}	s_{dx}	$s_{pdx,e}$
				Proportions of Variance			
Number of dimensions							
≤ 5 (k = 14, n = 7000)	20.89% ^{a}	3.59% ^{a}	**0.57%** ^{a}	15.16% ^{a}	20.89% ^{a}	**2.24%** ^{a}	36.64%
> 5 (k = 9, n = 4500)	11.43% ^{b}	4.69% ^{b}	**5.76%** ^{b}	9.33% ^{b}	25.61% ^{b}	**7.20%** ^{b}	38.02%
Fully crossed design?							
No (k = 11, n= 6000)	**13.07%** ^{a}	3.92% ^{a}	**4.89%** ^{a}	14.08% ^{a}	**15.71%** ^{a}	**6.61%** ^{a}	43.39%
Yes (k = 12, n= 5500)	**20.96%** ^{b}	4.12% ^{a}	**0.51%** ^{b}	11.78% ^{b}	**29.17%** ^{b}	**1.96%** ^{b}	31.48%
Behavioral checklists?							
Yes (k = 8, n= 3500)	**19.28%** ^{a}	**7.01%** ^{a}	**6.86%** ^{a}	11.39% ^{a}	21.23% ^{a}	5.25% ^{a}	30.77%
No (k = 15, n= 8000)	**16.27%** ^{b}	**2.72%** ^{b}	**0.07%** ^{b}	13.54% ^{b}	23.39% ^{b}	3.72% ^{b}	40.59%
AC purpose^{c}							
Selection (k = 9. n= 5500)	16.38% ^{a}	2.88% ^{a}	1.98% ^{a}	16.65% ^{a}	17.63% ^{a}	2.84% ^{a}	43.67%
Developmental (k = 11, n= 4500)	18.54% ^{b}	4.52% ^{b}	0.61% ^{b}	12.89% ^{b}	21.07% ^{b}	4.77% ^{b}	37.58%
Assessor occupation^{d}							
Non-Managers (k = 10, n= 5500)	19.53% ^{a}	2.97% ^{a}	0.61% ^{a}	13.55% ^{a}	23.48% ^{a}	3.93% ^{a}	34.82%
Managers (k = 10, n= 4500)	14.84% ^{b}	3.95% ^{b}	3.91% ^{b}	13.29% ^{a}	21.38% ^{b}	4.76% ^{b}	39.13%
				Cohen's D Effect Sizes			
Number of dimensions							
≤ 5 (k = 14, n = 7000)	.33	.20	**.99**	**.47**	.27	**1.37**	
> 5 (k = 9, n = 4500)	(.29, .37)	(.16, .24)	**(.96, 1.04)**	**(.43, .51)**	(.24, .31)	**(1.33, 1.41)**	
Fully crossed design?							
No (k = 11, n= 6000)	**.94**	.04	**.81**	.19	**.84**	**1.27**	
Yes (k = 12, n= 5500)	**(.91, .98)**	(.00, 07)	**(.77, .85)**	(.16, .23)	**(.80, .88)**	**(1.23, 1.30)**	
Behavioral checklists?							
Yes (k = 8, n= 3500)	**.95**	**.82**	**1.20**	.18	.12	.35	
No (k = 15, n= 8000)	**(.91, .98)**	**(.78, .86)**	**(1.15, 1.25)**	(.02, .14)	(.08, .16)	(.32, .40)	
AC purpose^{c}							
Selection (k = 9. n= 5500)	.23	.28	.36	.33	.23	**.47**	
Developmental (k = 11, n= 4500)	(.19, .27)	(.25, .32)	(.32, .40)	(.29, .36)	(.19, .27)	**(.43, .51)**	
Assessor occupation^{d}							
Non-Managers (k = 10, n= 5500)	**.50**	.18	**.58**	.02	.11	.19	
Managers (k = 10, n= 4500)	**(.46, .54)**	(.14, .22)	**(.54, .62)**	(.01, .06)	(.07, .15)	(.15, .23)	

Notes: ^{ab}Effect/design feature pairs with different superscripts are statistically different from one another at *p*< .05. ^{c}The purpose of the Bycio et al. (1987) AC did not fall exclusively into either category, thus it was not included in this analysis. ^{d}TheKolk et al. (2003) study did not identify the assessor occupation, thus it was not included in this analysis. k = numbers of studies with the characteristic in question. n = total sample size of Monte Carlo data sets with the characteristic in question. p = person; d = dimension; x = exercise; pd = person by dimension; px = person by exercise; dx = dimension by exercise; pdx,e = error. Effect sizes greater than .80 are in boldface type.

5.1.1 Number of Dimensions

A large effect was found for both the exercise and dimension by exercise effects (see Table 3). Assessment centers with five or fewer dimensions demonstrated a substantially lower exercise effect (*d* = .99) and a substantially higher person by dimension interaction (*d* = .47). However, no substantial differences emerged for the person, dimension, or person by exercise interaction effects. Thus, Hypothesis 1 was only partially supported. Furthermore, it should also be noted that the dimension by exercise interaction effect was larger when greater than five dimensions were utilized (*d* = 1.37). This suggests that, for these assessment centers, dimension observability is of greater concern when more than five dimensions are featured.

5.1.2 Crossed vs. Non-crossed Design

When comparing crossed and non-crossed designs, large effects were found for the person, exercise, and person by exerciseeffects. Both the person and person by exercise effects were substantially larger for fully crossed assessment centers (*d* = .94 and *d* = .84, respectively). In contrast, the exercise effects were smaller for those assessment centers that were fully crossed (*d* = .81).Thus Hypothesis 2 was only partially supported. A fully crossed design improved the construct-related validity via two sources of variance (i.e., the person and exercise effect) but reduced the construct-related validity for another source of variance (i.e., the person by exercise interaction). Additionally, similar to the finding concerning the number of dimensions examined by an assessment center, dimension observability is a greater concern for assessment centers that are not fully crossed (*d* = 1.27).

5.1.3 Behavioral Checklists

For the assessment centers analyzed, large effects were displayed for the person (*d* = .95), dimension (*d* = .82), and exercise effects (*d* = 1.20). Specifically, all three of these effects were larger when behavioral checklists were used. However, there were no substantial differences for the remaining effects. Thus, Hypothesis 3 was only partially supported. Although behavioral checklists improved the construct-related validity of assessment center ratings via two sources of variance (i.e., the person and dimension effects), the use of these checklists reduced the construct-related validity ratings via another source of variance (i.e., the exercise effect).

5.1.4 Assessment Center Purpose

Overall, none of the sources of variance demonstrated large effect sizes based on the purpose of the assessment center (selection vs. development). Thus, Hypothesis 4 was not supported. The purpose of the assessment center appears to be irrelevant to construct-related validity. Interestingly, developmental assessment centers demonstrated a substantially greater dimension by exercise interaction (*d* = .47).

5.1.5 Assessor Occupation

Moderate effects were displayed for the person and exercise effects. The person effect was substantially greater in the assessment centersin which non-managers were used as assessors in comparison with those assessment centers that utilized managers (*d* = .50). In contrast, the exercise effect was substantially smaller when non-managers were used as assessors (*d* = .58). Thus, Hypothesis 5 was only partially supported. Assessment centers that utilized professional assessors exhibited a significantly higher person effect and a significantly lower exercise effect. Otherwise, the occupation of the assessor had little impact on the construct-related validity of assessment center ratings.

6. Discussion

There have now been almost three decades of research on the construct-related validity of assessment center PEDRs. Much of this research has been less than promising and tends to diminish the effects of dimensions on the variance of PEDRs (e.g., Lance, Lambert et al., 2004). Primarily utilizing CFA techniques, research has shown that models with large exercise effects tend to fit PEDR data better than those with large dimension effects (e.g., Bycio et al., 1987; Fleenor, 1996;Lance, Foster, et al., 2004; Lance, Lambert, et al., 2004; Lance et al., 2000; Schneider & Schmidt, 1992). The present study sought to challenge these findings by utilizing a relatively overlooked technique, univariategeneralizability theory, to examine the sources of variance in assessment center PEDRs. Our results lead us to believe that the current view of assessment center construct-related validity is not as dire as some suggest (e.g., Lance, 2008) and may be better clarified with the utilization of generalizability theory. Specifically, although the person by exercise interaction was the largest single source of variance (21.83%), the person, dimension, and person by dimension effects together accounted for even more variance (34.06%). Thus, the very effects that are commonly considered indicators of construct-related validity (cf. Arthuret al., 2000; Bowler &Woehr, 2009) contributed more variance than the effects that are considered to be indicative of situational specificity (i.e., the exercise and person by exercise effects accounted for 24.43% of the variance).

Overall, the prevailing view that assessment centersdisplay either dimension effects or exercise effects is too simplistic. This is a rather broad generalization that does not provide much useful information for the improvement of assessment center design. With the utilization of generalizability theory, we now have a better language with which to discuss the issue. Our results suggest that although PEDRs are comprised of a substantial amount of variance associated with construct-related validity (i.e., person, dimension, and person by dimension effects), a substantial amount of situational specificity also exists in many assessment centers (i.e., person by exercise effects). However, both of these results were highly variable across assessment centers. Thus, the most parsimonious conclusion seems to be that some assessment centers are implemented better than others. Moreover, the prevailing view that assessment centers do not function in the manner in which they are designed is rather inappropriate. This conclusion is synonymous to examining the construct-related validity of numerous measures of cognitive ability and making general assertions regarding the nature of the construct based on single studies of the various measures. Simply stated, as there is no standard assessment center, there can be no prevailing view about the functionality of assessment centers as a whole. What is needed in the assessment center literature is more data on single assessment centers that employ slight variations from one implementation to another. This would allow for a more detailed understanding of the impact of particular design features on construct-related validity. This is turn would provide more beneficial guidelines for assessment center design and development.

6.1 Implications

The findings of this study have several significant implications for researchers and practitioners. The substantial amount of variance attributed to the person, dimension, and person by dimension effects reopens a discussion that was once stifled by the results generated by a questionable analytical technique. Instead of employing CFA and being confined to model fit, substantial evidence suggests that generalizability theory provides an excellent tool which will now allow researchers to explore other facets not readily examined in the current literature. For example, this study analyzed the person effect along with its various interactions and found a considerable amount of variance associated with these effects. Additionally, researchers may now explore additional facets, such as the aforementioned rater effect as well as exercise order effects, both of which could substantially affect the variance of PEDRs.

In addition, our results revealed large standard deviations for the person by exercise and person by dimension interaction effects. Both of these effects are crucialto identifying appropriate assessment center functionality. These significant standard deviations imply that the appropriate functioning of an assessment center is largely dependent on the implementation of the particular assessment center being evaluated. Although we have highlighted several design characteristics that can impact proper functioning (i.e., number of dimensions, crossed vs. non-crossed design, and use of behavioral checklists), future research should investigate the extent to which additional characteristics, such as exercise order and trait activation, can affect the magnitude of the proportion of variance attributed to either of the critical effects. Furthermore, based on our findings, in order to decrease the person by exercise effect and thus decrease situational specificity, administrators should focus on developing assessment centers that utilize a non-crossed design as well as reducing the number of dimensions measured in the assessment center.

6.2 Limitations

This study was not without its limitations. First, as with any meta-analysis, a major concern relates to obtaining enough data to make meaningful, supported statements. Although previous reviews have included over 30 MTMM matrices, our analyses necessitatedthe means and standard deviations of each variable in the MTMM matrix. This requirement limited the amount of data that was available for analysis. It would have been preferable to match the number of studies utilized by previous reviews such as Bowler and Woehr (2006) and Lance, Lambert et al. (2004). Unfortunately, this was not possible due to the necessary inclusion criteria. Second, it should be noted that the analyses included in this study were conducted on simulated data. Although it would have been optimal to analyze the original data from which the MTMM matrices were formed, that was neither feasible nor realistically possible. Therefore, 500 simulations were created and analyzed for each MTMM matrix. Even though this method surely produced some deviation from the original data, it was the most viable alternative for reassessing past assessment center data without utilizing a CFA. Finally, it should be noted that not all assessment center effects were included in our analyses. Previous studies have suggested that individual raters may influence assessment center PEDRs (e.g., Arthur et al., 2000; Kolk, Born, & Flier, 2002). Since multiple ratings are collapsed in the formation of MTMM matrices, the assessment of such an effect was not possible. Future research should further examine the rater effect, as well as its interactions with the other effects (e.g., person by rater, dimension by rater, exercise by rater), and its influence on assessment center PEDRs.

An additional limitation relates to the design features that were examined. Recent research has begun to look at additional features that impact the construct-related validity of PEDR. For example, Schollaert and Lievens (2011) noted that role player prompts appear to facilitate better dimension measurement. Similarly, Melchers, Kleinmann, and Prinz (2010) noted that having to simultaneously rate multiple individuals has a detrimental impact on ratings. Similarly, current work on task-based assessment centers has also shown some potential (e.g., Jackson, Stillman, &Englert, 2010) as has recent research into parallel forms of assessment center exercises

(Brummel, Rupp, & Spain, 2009). Unfortunately, features such as these, which have only begun to be studied, do not have any MTMM matrices that could be included in the analyses of this study.

7. Conclusions

As we have noted, the debate about assessment center construct-related validity has dominated the assessment center literature for more than 20 years. In its wake it has left most researchers and practitionersconvinced that assessment centers do not function in the way in which they are intended to function. Despite being designed to measure individual differences in performance dimensions (Bray & Grant, 1966), the prevailing view is that assessment centersactually measure some form of situational specificity. However, the little existing research that does not rely on a problematic statistical methodology (i.e., a CFA of an MTMM matrix) paints a much different picture of assessment center functioning. Although this research does not regard assessment centers as flawless, it does suggest that there are substantial dimension effects and that assessment centers bear considerable construct-related validity. Nonetheless, this research also highlights substantial person by exercise interactions. Thus, while supporting the construct-related validity of assessment center, strong support for exercise effects also exists. Using new methodologies, researchers must systematically identify the assessment center design features that alter this effect and attempt to minimize it. Furthermore, if this effect cannot be parsed out of assessment center functioning, at the very least it needs to be validated against job performance (e.g., Highhouse& Harris, 1993), as has been done with both assessment center overall assessment ratings and individual assessment center dimensions (Arthur et al., 2003).

References

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

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

*Becker, A. S. (1990). *The effects of a reduction in assessor roles on the convergent and discriminant validity of assessment center ratings.* Unpublished doctoral dissertation, University of Missouri, St. Louis.

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

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of inferential and evidential bases. *Journal of Applied Psychology*, *74,* 478-494. http://dx.doi.org/10.1037//0021-9010.74.3.478

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

Bowler, M. C., & Woehr, D. J. (2008, April). Evaluating assessment center construct-related validity via variance partitioning.In B. J. Hoffman (Chair), *Reexamining Assessment Centers: Alternate Approaches.* Paper presented at the 23rd annual meeting of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Bray, D. W., & Grant, D. L. (1966). The assessment center in measurement of potential for business management. *Psychological Monographs*, *80*(17), 1-27. http://dx.doi.org/10.1037/h0093895

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

Brummel, B. J., Rupp, D. E., & Spain, S. M. (2009). Constructing parallel simulation exercises for assessment centers and other forms of behavioral assessment. *Personnel Psychology, 62,* 137-170. http://dx.doi.org/10.1111/j.1744-6570.2008.01132.x

*Bycio, P., Alvares, K. M., & Hahn, J. (1987). Situational specificity in assessment center ratings: A confirmatory factor analysis. *Journal of Applied Psychology*, *72*, 463-474. http://dx.doi.org/10.1037/0021-9010.72.3.463

Campbell, D. T., & Fiske, D. W. (1959).Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105. http://dx.doi.org/10.1037/h0046016

Chan, D. (1996). Criterion and construct validation of an assessment centre. *The Journal of Occupational and Organizational Psychology, 69,* 167-181. http://dx.doi.org/10.1111/j.2044-8325.1996.tb00608.x

*Chorvat, V. P. (1994). *Toward the construct validity of assessment center leadership dimensions: A multitrait-multimethod investigation using confirmatory factor analysis.* Unpublished doctoral dissertation. University of South Florida.

Cochran, W. G., & Cox, G. M. (1957). *Experimental designs* (2nd ed.). New York, NY: John Wiley & Sons, Inc.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Conner, W. S., & Young, S. (September, 1961). *Fractional factorial designs for experiments with factors at two and three levels* (Applied Mathematics Series, 58). Washington, DC: National Bureau of Standards.

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

Donahue, L. M., Truxillo, D. M., Cornwell, J. M., & Gerrity, M. J. (1997). Assessment center validity and behavioral checklists: Some additional findings. *Journal of Social Behavior and Personality*, *12*, 85–108.

Eurich, T. L., Krause, D. E., Cigularov, K., & Thornton, G. C. (2009). Assessment centers: Current practices in the United States. *Journal of Business Psychology, 24,* 387-407. http://dx.doi.org/10.1007/s10869-009-9123-3

Federer, W. T. (1955). *Experimental design*. New York: Macmillan.

Fleenor, J. W. (1996). Constructs and developmental assessment centers: Further troubling empirical findings. *Journal of Business and Psychology, 10,* 319-333. http://dx.doi.org/10.1007/BF02249606

*Fredericks, A. J. (1989). *Assessment center ratings: Models and processes.* Unpublished doctoral dissertation, University of Nebraska, Lincoln.

Gaugler, B. B., & Thornton, G. C., III. (1989). Number of assessment center dimensions as a determinant of assessor generalizability of the assessment center ratings. *Journal of Applied Psychology, 74*, 611-618. http://dx.doi.org/10.1037/0021-9010.74.4.611

Hartley, H. O., Rao, J. N. K., & LaMotte, L. (1978). A simple synthesis-based method of variance component estimation. *Biometrics, 34,* 233-244. http://dx.doi.org/10.2307/2530013

Hemmerle, W. J., & Hartley, H. O. (1973). Computing maximum likelihood estimates for the mixed AOV model using the W-transformation. *Technometrics, 15,* 819-831. http://dx.doi.org/10.2307/1267392

Highhouse, S., & Harris, M. M. (1993). The measurement of assessment center situations: Bem's template matching technique for examining exercise similarity. *Journal of Applied Social Psychology, 23,* 1-55. http://dx.doi.org/10.1111/j.1559-1816.1993.tb01057.x

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

Jackson, D. J., Stillman, J. A., & Atkins, S. G. (2005). Rating task versus dimensions in assessment centers: A psychometric comparison. *Human Performance, 18,* 213-241. http://dx.doi.org/10.1207/s15327043hup1803_2

Jackson, D. J. R., Stillman, J. A., & Englert, P. (2010). Task-based assessment centers: Empirical support for a systems model. *International Journal of Selection and Assessment, 18,* 141-154. http://dx.doi.org/10.1111/j.1468-2389.2010.00496.x

Jansen, A., Lievens, F., & Kleinmann, M. (2011). Do individual differences in perceiving situational demands moderate the relationship between personality and assessment center ratings? *Human Performance, 24,* 231-250. http://dx.doi.org/10.1080/08959285.2011.580805

*Jansen, P. G. W., & Stoop, B. A. M. (2001). The dynamics of assessment center validity: results of a 7-year study. *Journal of Applied Psychology*, *86*, 741-753. http://dx.doi.org/10.1037/0021-9010.86.4.741

Joiner, D. A. (2002). Assessment centers: What's new? *Public Personnel Management, 31*(2), 179-185.

*Kolk, N. J., Born, M. P., & Flier, H. V. D. (2002). Impact of common rater variance on construct validity of assessment center dimension judgments. *Human Performance, 15,* 325-337. http://dx.doi.org/10.1207/S15327043HUP1504_02

*Kolk, N. J., Born, M. P., & Flier, H. V. D. (2003). The transparent assessment centre: The effects of revealing dimensions to candidates. *Applied Psychology: An International Review*, *52*, 648-668. http://dx.doi.org/10.1111/1464-0597.00156

*Kudisch, J. D., Ladd, R. T., & Dobbins, G. H. (1997). New evidence on the construct validity of diagnostic assessment centers: The findings might not be so troubling after all. *Journal of Social Behavior and Personality, 12,* 129-244.

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

*Lance, C. E., Foster, M. R., Gentry, W. A., & Thoresen, J. D. (2004). Assessor cognitive processes in an operational assessment center. *Journal of Applied Psychology*, *89*, 22-35. http://dx.doi.org/10.1037/0021-9010.89.1.22

Lance, C. E., Foster, M. R., Nemeth, Y. M., Gentry, W. A., & Drollinger, S. (2007). Extending the nomological network of assessment center construct validity: Prediction of cross-situationally consistent and specific aspects of assessment center performance. *Human Performance, 20,* 345-362. http://dx.doi.org/10.1207/S15327043HUP1304_1

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology*, *89*, 377-385. http://dx.doi.org/10.1037/0021-9010.89.2.377

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R., & Smith, D. E. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance*, *13*, 323-353. http://dx.doi.org/10.1207/S15327043HUP1304_1

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

Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology, 91,* 247-258. http://dx.doi.org/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 Psychology*, *86*, 1202-1222. http://dx.doi.org/10.1037/0021-9010.86.6.1202

*Lievens, F., Van Keer, E., Harris, M. M., & Bisqueret, C. (2003). Predicting cross-cultural training performance: the validity of personality, cognitive ability, and dimensions measured by an assessment center and a behavior description interview. *Journal of Applied Psychology, 88,* 476-489. http://dx.doi.org/10.1037/0021-9010.88.3.476

Melchers, K. G., Kleinmann, M., & Prinz, M. A. (2010). Do assessors have too much on their plates? The effects of simultaneously rating multiple assessment center candidates on rating quality. *International Journal of Selection and Assessment, 18,* 329-341. http://dx.doi.org/10.1111/j.1468-2389.2010.00516.x

*Parker, M. W. (1992). *A construct validation of the Florida Principal Competencies Assessment Center using confirmatory factor analysis.* Unpublished doctoral dissertation, University of South Florida, Tampa.

Petrides, K. V., Weinstein, Y., Chou, J., Furnham, A., & Swami, V. (2010). An investigation into assessment centre validity, fairness, and selection drivers. *Australian Journal of Psychology, 62,* 227-235. http://dx.doi.org/10.1080/00049531003667380

*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 Performance, 13,* 355-370. http://dx.doi.org/10.1207/S15327043HUP1304_2

*Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology*, *77*, 32-41. http://dx.doi.org/10.1037/0021-9010.77.1.32

Schollaert, E., & Lievens, F. (2011). The use of role-player prompts in assessment center exercises. *International Journal of Selection and Assessment, 19,* 190-197. http://dx.doi.org/10.1111/j.1468-2389.2011.00546.x

Spychalski, A. C., Quinones, M. A., Gaugler, B. B., & Pohley, K. (1999). A survey of assessment center practices in organizations in the United States. *Personnel Psychology, 50,* 71-90. http://dx.doi.org/10.1111/j.1744-6570.1997.tb00901.x

Thornton, G. C. (1992). *Assessment centers in human resource management.* Reading, MA: Addison-Wesley Publishing Company.

Thornton, G. C., & Mueller-Hanson, R. A. (2004). *Developing organizational simulations*. Philadelphia, PA: Lawrence Erlbaum Associates.

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

Woehr, D., & Feldman, J. (1993). Processing objective and question order effects on the causal relation between memory and judgment in performance appraisal: The tip of the iceberg. *Journal of Applied Psychology*, *78,* 232-241. http://dx.doi.org/10.1037/0021-9010.78.2.232