EVIDENCE OF CONSTRUCT-RELATED VALIDITY FOR ASSESSMENT CENTERS: MORE PIECES OF THE INFERENTIAL PIE

A Dissertation

by

KATHRYN DIANE ARCHULETA

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2009

Major Subject: Psychology

EVIDENCE OF CONSTRUCT-RELATED VALIDITY FOR ASSESSMENT CENTERS: MORE PIECES OF THE INFERENTIAL PIE

A Dissertation

by

KATHRYN DIANE ARCHULETA

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Winfred Arthur, Jr.
Committee Members,	Mindy Bergman
	David J. Martin
	Stephanie Payne
Head of Department,	Les Morey

May 2009

Major Subject: Psychology

ABSTRACT

Evidence of Construct-Related Validity for Assessment Centers: More Pieces of the Inferential Pie. (May 2009)

Kathryn Diane Archuleta, B.A., Rice University; M.S., Texas A&M University

Chair of Advisory Committee: Dr. Winfred Arthur, Jr.

Much research has been conducted on the topic of the construct-related validity of assessment centers, however a definitive conclusion has yet to be drawn. The central question of this debate is whether assessment centers are measuring the dimensions they are designed to measure. The present study attempted to provide more evidence toward the improvement of construct-related validity. The first hypothesis involved determining whether opportunity to observe and opportunity to behave influenced discriminant and convergent validity. The second hypothesis addressed the debate over evaluation method and examined which method, within-exercise or within-dimension, yielded more favorable internal construct-related validity evidence. The third hypothesis explored the call for exercise scoring in assessment centers and compared the criterion-related validity of exercise versus dimension scores within the same assessment center. Finally, the fourth objective looked at the relationship of the stability of the dimensions with internal construct-related validity, specifically convergent validity evidence. A developmental assessment center used in two applied settings supplied the data. Two administrations of the assessment center were conducted for low to mid-level managers

in a state agency (N = 31). Five administrations were conducted in a professional graduate school of public administration that prepares students for leadership and managerial positions in government and public service (N = 108). The seven administrations yielded a total sample size of 139 participants.

Analysis of multi-trait-multi-method (MTMM) matrices revealed that, as hypothesized, a lack of opportunity to behave within exercises, operationalized using behavior counts, yielded poor discriminant validity. Assessor ratings of opportunity to observe and behave did not produce hypothesized results. Consistent with the second hypothesis, secondary assessors, who represented the within-dimension evaluation method, provided ratings that demonstrated better construct-related validity evidence than the ratings provided by primary assessors, who represented the within-exercise method. Correlation and regression analyses of the dimension/performance relationships and the exercise/performance relationships revealed neither dimensions nor exercises to be the better predictor of supervisor ratings of performance. Using MTMM, partial support was found for the fourth objective: those dimensions that were more stable across exercises yielded better convergent validity evidence versus those dimensions that were more situationally specific. However the differences were not statistically significant or large. Overall results of this study suggest that there *are* some areas of design and implementation that can affect the construct-related validity of assessment centers, and researchers should continue to search for ways to improve assessment center construct-related validity, but should also look for ways other than MTMM to assess validity.

DEDICATION

This is dedicated to my son, Max, who lost countless hours of Mommy time while I worked on it.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Winfred Arthur, Jr., for getting me started and then sticking by me to the end, and my committee members, Dr. Bergman, Dr. Martin, and Dr. Payne, for their guidance and support throughout the course of this research.

Thanks also go to my mom (and my sons' Granny) for the trips to College Station with both of my kids when they were infants, and for the many visits to babysit while I worked. A thank you to Nancy Lane as well for the "dissertation duty" babysitting help. I honestly could not have finished this without the two of you doing what you did.

I would like to say a big thank you to Anton Villado and Meg Horner who served as my "signature getters" while I was in Houston. You were always there for me and never missed a deadline – even when I emailed you last minute. I also want to extend my gratitude to the research assistants who helped code the data.

Finally, thanks to my husband who has supported me in this long endeavor and supported the family while I worked to complete this dissertation. I know it's been hard, Jay, and I really do appreciate your patience. And to my son, Arch, who held off on being born until I had completed my first draft of the final document – thank you for that!

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

INTRODUCTION

The purpose of this dissertation is to investigate the construct-related validity of assessment centers. Research has generally shown a lack of construct-related validity evidence for assessment centers; however, some studies have shown otherwise. Many reasons for this discrepancy have been posited and will be reviewed within the context of this dissertation. No research line within this area has provided a conclusive answer (e.g., see focal article "Why assessment centers do not work the way they are supposed to" and commentaries in *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2008, 1(1)*). The present study attempts to provide new ways of examining old theories, as well as present an exploratory area that has not been investigated. Each of the proposed research questions will provide another piece of evidence that pertains to the debate over the construct-related validity of assessment centers.

An assessment center is a method of collecting information about a participant, just as an interview is designed to collect information; and just as interviews are not all developed and implemented in the same manner, neither are assessment centers. The basic premise of an assessment center is that it is a method that uses multiple techniques, with at least one job-related simulation, multiple assessors, and a pooling of information or data (Joiner, 2000; Reilly, Henry, & Smither, 1990). However, what specific techniques and simulations are used, what dimensions are measured, and how the

---

This dissertation follows the style and format of the *Journal of Applied Psychology*.

assessment center is designed and implemented can be adapted for individual assessment center use. For example, the method of evaluation can vary across assessment centers - some assessment centers have assessors rate after each exercise, others have assessors wait until the conclusion to make any ratings (Sackett & Dreher, 1982); the number of dimensions rated can range from 3 to 25 (Woehr & Arthur, 2003). These variations make the assessment center a constantly changing and evolving method, and potentially more difficult to study. Regardless of these differences and the complexity of research surrounding the method, assessment centers continue to be used for selection and promotion decisions, as well as for developmental purposes (e.g., Fitzgerald & Quaintance, 1982; Howard, 1997).

Validity of Assessment Centers

Although the debate over what an assessment center is has been addressed with the publication of the Guidelines for Assessment Center Operations (Joiner, 2000), the debate over the validity of assessment centers has yet to reach a satisfactory conclusion. Past research has generally held that assessment centers demonstrate criterion-related validity with a lack of construct-related validity.

Criterion-Related Validity of Assessment Centers

One of the most consistent findings in the literature is criterion-related validity evidence for assessment centers (Schleicher, Day, Mayes, & Riggio, 1999). Since the method's inception, studies have consistently shown that assessment centers demonstrate acceptable levels of criterion-related validity with respect to training outcomes, performance ratings, potential, career advancement, and various other criteria (e.g.,

Borman, 1982; Chan, 1996; McEvoy & Beatty, 1989; McEvoy, Beatty, & Bernardin, 1987; Turnage & Muchinsky, 1984; Tziner & Dolan, 1982). Meta-analyses have echoed these results (Gaugler, Rosenthal, Thornton, & Bentson, 1987; Schmitt, Gooding, Noe, & Kirsch, 1984). Gaugler et al. (1987) found an overall validity coefficient of .37 (corrected) for assessment centers, and argued for both validity generalization and situation specificity. They found that the criterion-related validity of assessment centers did generalize across situations, but also found that some variance is due to different implementations and designs of assessment centers. Moderators identified by Gaugler et al. included the number of exercises, such that a higher number of different types of exercises yielded higher validity, and type of assessor, where psychologists yielded higher validity coefficients versus managers.

One drawback of this meta-analysis is that it did not break down the studies by dimension score, but looked only at overall assessment ratings (OARs). The argument could be made that some dimensions have higher validities than others do. In fact, a meta-analysis by Arthur, Day, McNelly, and Edens (2003) found that four individual dimensions (problem solving, influencing others, organizing and planning, and communication) demonstrated validity coefficients equal to or higher than the .37 found by Gaugler et al. Further, these four dimensions, when treated as unique contributors, accounted for more variance in the prediction of performance than did the OAR assessed by Gaugler et al. (20% versus 14%). This more recent meta-analysis highlights the need of assessment center research to move away from solely examining the OAR, and to

increase examination of dimension- and exercise-level information (see also Edwards, 2001).

In addition to demonstrating overall predictive validity, assessment centers have shown incremental validity over other traditional predictors (e.g., Chan, 1996; Krause, Kersting, Heggestad, & Thornton, 2006; McEvoy & Beatty, 1989; McEvoy et al., 1987). For example, in a meta-analytic review of the relationship of assessment center dimensions to the Big Five personality traits and cognitive ability, Meriac, Hoffman, Woehr, and Fleisher (2008) found that assessment center dimensions shared only a small amount of variance with either cognitive ability or personality. It was also found that assessment center dimensions added significant unique variance (change in R^2) to the prediction job performance, above the variance accounted for by personality and cognitive ability. The criterion-related validity of assessment centers has been further demonstrated with longitudinal data, showing predictive evidence for managerial level (Ritchie & Moses, 1983) and salary growth (Jansen & Stoop, 2001) seven years after implementation. McEvoy and Beatty (1989) concluded that assessment centers have unique long-term value for predicting future performance, and can therefore be considered a cost-effective method.

There is also some evidence that assessment centers show little to no adverse impact or differential validity (Baron & Janman, 1996; Dean, Roth, & Bobko, 2008; Norton, 1977; Pynes & Bernardin, 1992), although Dean et al. (2008) found relatively large subgroup differences for Black-White comparisons and cautioned that adverse impact for this group may be worse than previously reported. Finally, it is found that

applicants view assessment centers as "more face valid, acceptable, and fair than paper-and-pencil tests" (Howard, 1997, p.18; see also Macan, Avedon, Paese, & Smith, 1994).

It should be noted that most research on assessment centers investigates assessment centers as a method independent of the constructs being measured. This oversimplification complicates the study of assessment centers, but as with the interview and other methods, it does not necessarily mean that research should not be conducted. Rather, constructs being measured should be taken into consideration when examining results from studies on assessment centers as a method. The constructs play as important a role in validity and utility as does the method used to measure those constructs. All research using assessment centers as a methodology provides results that are confounded by the dimensions being measured (see Arthur & Villado, 2008, for discussion of construct-method distinction for predictors in general). The present study attempts to provide some generalizations about assessment centers as a method, but does not ignore the fact that constructs measured within an assessment center can impact results.

In sum, research shows that assessment centers demonstrate criterion-related validity evidence across settings and across different criteria, and that they provide incremental and unique validity above traditional predictors. Therefore, assessment centers have a place in making selection and promotion decisions.

Content-Related Validity of Assessment Centers

Although the criterion-related validity of assessment centers has been consistently demonstrated through various primary studies and meta-analyses, the question of the content-related validity of assessment centers has not received nearly as

much published attention (Hoeft & Schuler, 2001). A review of the assessment center literature revealed no published study that actually documented content-related validity evidence. Articles did document how to establish content-related validity for assessment centers (Adler, 1987; Schmitt & Noe, 1983) and how to develop assessment centers to demonstrate content-related validity (Ahmed, Payne, & Whiddett, 1997; Dulewicz, 1991; Sackett, 1987), but no study provided actual content-related validity evidence of an assessment center. Indeed, the more general finding was the statement that assessment centers have content-related validity because it is generally assumed that assessment centers, by definition, demonstrate content-related validity (Woehr & Arthur, 2003). Sackett (1982) sums up the issue by saying that "a belief that content validity is inherent in assessment centers is an over-simplification of a complex issue" (p. 143).

One aspect of the content-related validity of assessment centers that *has* been adequately documented is the importance of a thorough job analysis and careful attention paid to exercise construction when establishing this type of validity evidence (Adler, 1987; Dreher & Sackett, 1981; Joiner, 2000; Neidig, Martin, & Yates, 1979; Norton, 1977; Sackett, 1987; Schmitt & Noe, 1983; Schmitt & Ostroff, 1986; Tziner & Dolan, 1982). However, there is some debate over whether a content-related validation strategy is even appropriate for determining the validity of assessment centers (Dreher & Sackett, 1981; Sackett, 1987; Sackett & Dreher, 1982). The confusion seems to come from what factors are included in establishing content-related validity. Binning and Barrett's (1989) definition of content-related validity evidence focuses on the extent of overlap between the predictor measure and the performance domain of interest.

Operationally, this means determining whether the predictor is an adequate and appropriate sample of what a person might actually do on the job. The basic logic is, "if an applicant performs behaviors as part of the assessment phase that closely resemble behaviors in the performance domain" (Binning & Barrett, 1989, p. 482), then this is evidence for content-related validity. The confusion in the literature for assessment center content-related validity comes from whether the behaviors performed in the assessment phase should be examined at the exercise or dimension level, or whether both should be included.

Some researchers question the use of a content-related validation strategy for assessment centers (Dreher & Sackett, 1981; Sackett & Dreher, 1981, 1982). They argue that because dimension ratings, and not exercise ratings, are the outcome of assessment centers, validation should be based on dimensions, and a content-related validity approach is not appropriate for trait-based dimensions. (Note that not all assessment center dimensions are defined as being "trait-based;" many are defined as "behavior-based.") Further, they argue that a content-related validation strategy based on dimensions is inadequate for providing the sole evidence of the job-relatedness of assessment centers because the dimensions typically measured in assessment centers are too complex and lack the relevance necessary for that strategy. In response, some researchers have emphasized the use of exercises in determining content-related validity (Crawley, Pinder, & Herriot, 1990; Neidig & Neidig, 1984) stating that the multiple exercises in an assessment center are included to "adequately sample the relevant content domain" (Neidig & Neidig, 1984, p. 183). Although both sides provide valid points,

many authors, including the one of the present study, have realized the need to assess content-related validity based on the entire assessment center process, and not just exercises or dimensions (Ahmed et al., 1997; Norton, 1977, 1981; Tenopyr, 1977).

When viewed through the Binning and Barrett (1989) definition of content-related validity, which emphasizes *behavior* overlap between the performance and assessment domains, it seems relevant to include both dimensions and exercises in determining validity. Exercises are important because they provide the means of eliciting behaviors - the opportunity to perform; dimensions are important because they represent clusters of behaviors. It should be noted, however, that not all assessment centers use dimension definitions that have behaviors as indicators and that some dimension definitions are more closely related to trait descriptions than behaviors. "Trait descriptions" within this context generally refers to the tone of the definition and how the information will be used. In other words, are inferences made about a person's personality or are behaviors taken at face value? Further, trait descriptions often include the phrase "ability to" in front of behavior descriptions, such as "ability to break a problem into its essential parts" (Silverman, Dalessio, Woods, & Johnson, 1986). This again highlights the need to carefully consider the dimensions being measured when interpreting assessment center research.

Within the current view of validity, the once believed notion of validity as "types" has given way to the now accepted belief that all validation research provides evidence of the overall construct validity of a measure. The "tripartite" view of validity (content, criterion, and construct viewed as three different types of validity) led to

arguments about what type of validity was most appropriate for what test, and fueled the exercise/dimension debate of the content-related validity of assessment centers. The "unitarian" view of validity (content-, criterion-, and construct-related as examples of approaches to providing validity evidence) views validation efforts as ways to judgmentally or empirically provide support for inferences (Binning & Barrett, 1989; Society for Industrial and Organizational Psychology, Inc., 2003). From a personnel standpoint, these inferences are paths from the job performance domain to the predictor test; and validity evidence, regardless of the approach (content-, criterion-, or construct-related), provides support for this overall inference (construct validity). The approach is less important than the act of accumulating evidence.

Therefore, instead of focusing on the particular validation strategy used, research should concentrate on providing adequate validation evidence, in that "the number of validity analyses available is limited only by the creativity and experience of the analyst" (Landy, 1986, p. 1186). In light of this, the distinction between exercises and dimensions as they relate to content-related validity is less relevant (Schleicher et al., 1999). Indeed, the present study attempts to provide more pieces of evidence towards the overall construct validity of assessment centers.

Construct-Related Validity of Assessment Centers

The most prolific debate in the assessment center literature comes from studies examining the construct-related validity evidence of the method. The general finding has been the so-called exercise effect of assessment centers, demonstrated by moderate convergent validity evidence and little to no discriminant/divergent validity evidence

(e.g., Bycio, Alvares, & Hahn, 1987; Klimoski & Brickner, 1987; Neidig et al., 1979; Sackett & Dreher, 1982; Turnage & Muchinsky, 1982). Specifically, studies report empirically larger exercise factors versus dimension factors, and theoretically call into question the underlying factor structure of assessment center ratings. In other words, the interpretation of the results has been that it is the exercises that drive assessors' ratings, not the dimensions.

Internal Construct-Related Validity

The internal construct-related validity of assessment centers historically has been examined by using either multi-trait-multi-method (MTMM) matrices or factor analysis. When using the MTMM method to assess the internal construct-related validity of assessment centers, evidence for convergent validity is demonstrated with relatively large within-dimension, across-exercise (monotrait-heteromethod) correlations indicating that the respective dimension is measured across the different exercises. Evidence for discriminant validity is found with relatively small across-dimension, within-exercise (heterotrait-monomethod) correlations indicating that the different dimensions are not being measured based on the performance in the specific exercise, but are indeed distinct constructs. For assessment centers, however, this pattern has not emerged. Instead most studies have reported average heterotrait-monomethod (HTMM) correlations that are larger than average monotrait-heteromethod (MTHM) correlations, leading researchers to conclude a lack of construct-related validity for assessment centers (e.g., Bycio et al., 1987; Chan, 1996; Crawley et al., 1990; Jansen & Stoop, 2001; Joyce, Thayer, & Pond, 1994; Kudisch, Ladd, & Dobbins, 1997; Reilly et al.,

1990; Robertson, Gratton, & Sharpley, 1987; Russell, 1987; Silverman, Dalessio, Woods, & Johnson, 1986). In a review of MTMM assessment center studies, Jones (1992) reported an average MTHM correlation of .39, indicating a moderate level of convergent validity; however, an average HTMM correlation of .58 was found, indicating a lack of discriminant validity. This result has been so robust that many researchers have called for a shift from dimension scoring to exercise, or "role," scores (e.g., Hoeft & Schulet, 2001; Klimoski & Brickner, 1987; Lance, 2008b; Lance, Newbolt, Gatewood, Foster, French, & Smith, 2000; McEvoy et al., 1987; Robertson et al., 1987; Russell & Domm, 1995; Sackett & Dreher, 1984; Sackett & Harris, 1988; Silverman et al., 1986).

The first issue of *Industrial and Organizational Psychology: Perspectives on Science and Practice* (2008) was dedicated to the question of why assessment centers do not demonstrate construct-related validity. One of the conclusions of this group of articles was that because using MTMM implies that assessment center dimensions are traits, MTMM is not the best approach for assessing the construct-related validity of assessment centers and that using this method has led much of the assessment center research down a non-productive path. Thus, using MTMM in any further studies of the internal construct-related validity of assessment centers is “not encourage[d]” (Lance, 2008a, p. 144). However, in order to make comparisons to what research has found in the past, it is necessary within the present study to continue to use MTMM as the method of assessing construct-related validity. If the methodology is changed as well as the construct of interest, direct comparisons to past research cannot be made, and

improvements in validity cannot properly be assessed. Therefore, although MTMM is a suboptimal approach to assessing the internal construct-related validity of assessment centers, it is nonetheless what will be used for aspects of this study in order to allow for comparisons to past research.

Using factor analysis to assess the internal construct-related validity of ACs has yielded similar conclusions to using MTMM (e.g., Bycio et al., 1987; Chan, 1996; Fleenor, 1996; Highhouse & Harris, 1993; Joyce et al., 1994; Robertson et al., 1987; Russell, 1985; Sackett & Dreher, 1982). For example, Sackett and Dreher (1982) found that the underlying factor structure of assessment centers represented exercises not dimensions for two of the organizations they examined. Bycio et al. (1987) found that, although the confirmatory factor analysis did not produce a clear answer, when looking at the factor loadings the exercise loadings were higher and accounted for more variance than did the dimension factor loadings. These authors concluded that ratings made in this assessment center were largely situation, or exercise, specific. A review of assessment center factor analysis studies revealed that the factors that emerge are usually less than the number of dimensions rated and are usually interpreted as exercise factors (Jones, 1992). Thus, factor analysis results also have led some to question the construct-related validity of assessment centers.

A small number of studies have examined the construct-related validity of assessment centers by partitioning variance. The general finding is, again, an exercise effect with lower variance being accounted for by dimensions versus exercises (Schneider & Schmitt, 1992; Silverman et al., 1986; Turnage & Muchinsky, 1982).

Turnage and Muchinsky (1982) found a Person by Exercise interaction, indicating strong situation effects, coupled with a weak Person by Dimension interaction, indicating that assessors were not ordering candidates differently based on the dimensions. Although these studies did not account for all variance components, they do provide more evidence for a lack of construct-related validity for assessment centers.

External Construct-Related Validity

Because overall, not within-exercise, dimension scores are generally used for feedback in developmental assessment centers and for prediction in selection and promotion studies, some researchers have turned to the nomological network approach to provide construct-related validity evidence for assessment centers. This approach uses MTMM to assess the relationship of overall (across-exercise) dimension scores to other methods (outside of the assessment center ratings) that measure similar constructs. Thus, this method uses overall dimension scores only and does not consider the individual within-exercise dimension scores that are used in internal MTMM and most factor analytic studies of assessment centers. Unfortunately, most of the nomological network studies have not found strong support for the construct-related validity of assessment centers (Chan, 1996; Crawley et al., 1990; Fleenor, 1996; McEvoy et al., 1987; Shore, Shore, & Thornton, 1992). For example, Chan (1996) compared assessment center dimension ratings to paper-and-pencil tests of cognitive ability, cognitive-style, and personality. The results indicated no significant differences between conceptually related and conceptually unrelated relationships, and Chan concluded that

the "nomological network approach indicated no evidence of external construct validity" (p. 175).

The construct-related validity of assessment centers has come under attack because of the above-cited research. Studies using MTMM matrices and factor analysis to examine internal construct-related validity generally have found exercise factors instead of dimension factors. Studies assessing the variance components have found exercises to account for more variance. Studies examining the external construct-related validity have found little support. These studies all question the use of assessment centers, especially in a developmental arena.

Some Evidence of Construct-Related Validity

Although the preponderance of evidence points to a lack of construct-related validity for assessment centers, there are some studies that have found evidence of internal (Arthur, Woehr, & Maldegen, 2000; Kleinmann & Koller, 1997; Kudisch et al., 1997; Lievens & Keer, 1999; Sackett & Harris, 1988), as well as external (Shore, Thornton, & Shore, 1990; Thornton, Tziner, Dahan, Clevenger, & Meir, 1997) construct-related validity for assessment centers. For example, Arthur et al. (2000) found an average MTHM correlation greater than the average HTMM correlation (.60 versus .39). Further, one confirmatory factor analysis showed that a model with six dimensions and four exercises had the best fit, with further evidence of discriminant validity demonstrated by low dimension intercorrelations (Kudisch et al., 1997). Another study found significantly higher correlations for dimension ratings with comparable versus with noncomparable tests, indicating evidence of external

construct-related validity (Thornton et al., 1997). These studies, although small in number, question the received doctrine that assessment centers lack construct-related validity.

Why Does Research Generally Show a Lack of Construct-Related Validity for Assessment Centers?

Many researchers have theorized about and examined the reasons why a majority of studies have shown a lack of construct-related validity for assessment centers with a small handful of studies showing otherwise. The following sections discuss a variety of possible factors that may lead to finding versus not finding construct-related validity evidence in assessment centers. They are grouped into three main categories: design and implementation differences, problems with statistical approaches used, and construct misspecification.

Design and Implementation

Many studies of assessment centers either implicitly or explicitly theorize that the reason that some assessment centers demonstrate construct-related validity while others do not is because of methodological differences in design and implementation (Arthur et al., 2000; Dulewicz, 1991; Haaland & Christiansen, 2002; Howard, 1997; Jones, 1992; Kleinmann & Koller, 1997; Kuptsch, Kleinmann, & Koller, 1998; Lievens, 1998, 2002; Lievens & Conway, 2001; Schmitt, Schneider, & Cohen, 1990; Woehr & Arthur, 2003): studies that find low construct-related validity within an assessment center have flaws in the assessment center design and/or implementation. In other words, there is measurement error. Many factors have been hypothesized under the umbrella of

methodological differences in design and implementation. These factors can be categorized as follows: assessors (type and training), cognitive or information processing demands of assessors (observation and information gathering process, dimension definitions, number of dimensions, and ability of assessors), exercise and dimension design (opportunity to behave and observe), and evaluation approach.

Type of Assessor

One reason posited for the disparity of construct-related validity results is the type of assessor used in the assessment center. There are two types of assessors generally used, psychologists and managers, and although these groups often are used independently, some assessment centers use assessment teams made up of both psychologists and managers. Psychologists are considered more objective than managers, bringing to the assessment center only that which is learned in training with no pre-conceived notions, stereotypes, or biases regarding the specific participants or job (Adler, 1987). Managers, on the other hand, may use exercises as cues to the participants' on-the-job behaviors as opposed to rating participants based on assessment center behavior alone, or may provide ratings of dimensions that they feel are important but have, for whatever reason, been left out of the assessment center process (Jones, 1992; Lievens, 2001b). This has led some to argue that psychologists should play a key role as assessors in assessment centers (Lievens, 1998).

Gaugler et al.'s (1987) meta-analysis provided empirical evidence to support this idea. Type of assessor was found to be a moderator of criterion-related validity of assessment centers, with psychologists yielding higher validity coefficients than

managers. From this it is not a large leap to question the impact of type of assessor on the construct-related validity of assessment centers. In fact, two studies empirically tested this idea with expected results (Lievens & Conway, 2001; Sagie & Magnezy, 1997). In the Sagie and Magnezy (1997) study, ratings for five dimensions were subjected to factor analysis. The best-fit model for psychologists was a five-factor model fitting the dimensions, while the best fit model for managers was a two-factor model, which collapsed the five dimensions into two categories. The Lievens and Conway (2001) study asked the question, "Which design characteristics increase dimension variance?" and found that, when compared to managers, psychologists' ratings led to significantly higher proportions of dimension variance versus exercise variance. Further, a recent meta-analysis of design characteristics (Woehr & Arthur, 2003) found that using psychologists as assessors yielded more positive construct-related validity evidence (both convergent and divergent) than using managers. Thus, the hypothesis that type of assessor impacts the construct-related validity of assessment centers has received much support, and should be taken into consideration when designing and implementing an assessment center.

Training of Assessors

One offered explanation for the lack of construct-related validity in assessment centers involves the length and type of assessor training (Dugan, 1988; Dulewicz, 1991; Kauffman, Jex, Love, & Libkuman, 1993; Lievens, 1998, 2001a, 2002; Lievens & Conway, 2001; Schleischer et al., 1999; Woehr & Arthur, 2003). Although theoretically it seems that longer training should yield better validity results, study results have not

always supported this assumption. Dugan (1988), faced with research that states that assessors do not use all dimension information to formulate their OARs (Neidig et al., 1979; Russell, 1985), examined whether length of training would affect how much information was used to make an OAR. Results were inconsistent with expectations and showed that longer training did not lead assessors to use more dimensions in making their OARs.

Summary empirical examinations of the effect of length of training on the construct-related validity of assessment centers have provided equivocal results. In their review of 34 MTMM studies, Lievens and Conway (2001) concluded that shorter training (1 day or less) leads to significantly higher proportions of dimension variance versus longer training (more than 1 day). On the other hand, Woehr and Arthur's (2003) meta-analytic review of 31 MTMM studies indicated that longer training leads to more positive construct-related validity evidence. Although the issue of whether training should be longer or shorter is not resolved, it is plausible that length of training can affect the construct-related validity of assessment centers, thus potentially accounting for some of the disparity in results.

Researchers have begun to investigate the impact of *type* of training on the construct-related validity of assessment centers. Borrowing from the performance appraisal literature, frame-of-reference (FOR) training has been suggested as not only a reason why there are differences found in construct-related validity, but also as a method for improving convergent and discriminant validity. Two separate lab studies comparing FOR training to a control training group found that assessors who went through FOR

training were better at discriminating between dimensions, demonstrated higher external validity, had higher interrater reliability, and were more accurate (Lievens, 2001a; Schleicher et al., 1999). Although FOR was the focus of these studies, Woehr and Arthur (2003) concluded that what is most important is having training at all. This sentiment is echoed in the meta-analysis that showed that merely training assessors versus not training them produced stronger construct-related validity evidence.

Cognitive and Information Processing Demands

One of the most cited potential reasons for the exercise effect in assessment centers is the limited information processing capabilities of assessors coupled with the large cognitive demands placed on them during the assessment center process (Arthur et al., 2000; Bycio et al., 1987; Donahue, Truxillo, Cornwell, & Gerrity, 1997; Dugan, 1988; Gaugler & Thornton, 1989; Hoeft & Schuler, 2001; Jones, 1992; Kleinmann & Koller, 1997; Lievens, 1998, 2001b, 2002; Lievens & Conway, 2001; Reilly et al., 1990; Russell, 1985; Schleischer et al., 1999). In assessment centers, assessors are required to watch and record behaviors, classify behaviors into dimensions, rate the behaviors by dimension, and often reach consensus regarding overall ratings. Sometimes assessors are participants in the exercises, serving as role players or interviewers. Often, assessors are watching more than one candidate per exercise, and often the assessment process lasts eight hours, sometimes for more than one day. All of these conditions can potentially tax cognitive capacity by placing high demands on the assessors.

One reason this explanation has received attention is the group of studies that indicate that assessors are basing their OARs on only a few of the rated dimensions

(Dugan, 1988; Fletcher & Dulewicz, 1984; Gaugler & Thornton, 1989; Neidig et al., 1979; Russell, 1985; Turnage & Muchinsky, 1982). For example, Russell (1985) found that assessors' OARs were dominated by a single category instead of all 16 dimensions assessed, and Neidig et al. (1979) found that only five of 19 rated dimensions contributed significant unique variance to the overall score.

The explanation that high cognitive load has led to low construct-related validity for assessment centers has been addressed a number of ways. Many have attempted to reduce the cognitive load - through the definitions and number of dimensions, or through assessor training (which was addressed in the previous section). Some have questioned whether assessors, as humans, are even able to appropriately accomplish the task of assessor.

Cognitive and Information Processing Demands - Dimension Definitions

Poorly defined dimensions, ones that are not behaviorally or operationally focused, lead to increased cognitive demands on assessors (Dulewicz, 1991; Fitzgerald & Quaintance, 1982; Howard, 1997; Jones, 1992; Joyce et al., 1994; Lievens & Conway, 2001; Reilly et al., 1990; Turnage & Muchinsky, 1982). Assessors may have to make decisions about what behaviors fall under which dimension, and with each decision comes more subjectivity and lowered construct-related validity. Further, there is sometimes overlap in dimension definitions - some behaviors occur under more than one dimension (Brannick, Michaels, & Baker, 1989; Highhouse & Harris, 1993; Reilly et al., 1990; Schleicher et al., 1999; Turnage & Muchinsky, 1984). This can lead to increased HTMM correlations. Often, dimension definitions vary from exercise to exercise, even

if only slightly, potentially causing assessors to redefine the dimension to be exercise-specific (Ahmed et al., 1997; Joyce et al., 1994; Kauffman et al., 1993; Robertson et al., 1987; Schleicher et al., 1999). This could lead to lowered MTHM correlations and raised HTMM correlations. Together, these problems with dimension definitions can lead to poor construct-related validity results for assessment centers.

Some research has attempted to address the problem of dimension definitions through behavioral checklists and frame-of-reference (FOR) training. Although the two studies that explicitly mentioned using either behavioral checklists (Reilly et al., 1990) or FOR training (Schleischer et al., 1999) to provide better dimension definitions both demonstrated positive results for convergent and divergent validity respectively, no study has directly looked at the empirical impact of dimension definitions on the construct-related validity of assessment centers. However, in a qualitative review of moderators of construct-related validity of assessment centers, Lievens (1998) concluded that making dimensions conceptually distinct helps to increase discriminant validity. And in a survey of best practices, Ahmed et al. (1997) recommended using one common rating scale throughout the assessment center to reduce the cognitive load of assessors and thus increase construct-related validity. These two studies provide some evidence that dimension definitions affect the construct-related validity of assessment centers. Clearly, more research in this area needs to be conducted.

One possibility mentioned for the high level of dimension intercorrelations found (especially in factor analysis studies) is that perhaps dimensions are intended, by design, to be somewhat correlated (Jones, 1992; Robertson et al., 1987). In fact, Arthur et al.

(2003), in an investigation of the criterion-related validity of individual assessment center dimensions, found that it is unlikely that assessment center dimensions are independent of each other. With respect to construct-related validity, the relationship of the dimensions is reflected in the discriminant validity coefficients. Therefore, any true relationship of dimensions will artificially inflate the HTMM correlations resulting in potentially negative discriminant validity results. In sum, although much of the research with respect to dimension definitions has been qualitative or theoretical in nature, it seems reasonable for future research to address the impact that dimension definitions can have on the construct-related validity of assessment centers.

Cognitive and Information Processing Demands - Number of Dimensions

Buoyed by the group of studies that found that assessors are often collapsing their ratings into a smaller number of global ratings, the number of dimensions rated has received some attention as a moderator of the construct-related validity of assessment centers. A large number of dimensions may increase the complexity of the assessor's job, thus increasing the cognitive load of assessors and decreasing construct-related validity. One solution is to reduce the number of dimensions rated in each exercise in order to reduce the cognitive demands placed on assessors, which should in turn increase construct-related validity (Gaugler & Thornton, 1989; Kleinmann & Koller, 1997; Lievens, 1998, 2002; Lievens & Conway, 2000; Schneider & Schmitt, 1992; Turnage & Muchinsky, 1982; Woehr & Arthur, 2003). Reducing the number of dimensions also could result in making the dimensions more distinct and reducing the level of behavior overlap between dimensions (Gaugler & Thornton, 1989; Lievens & Conway, 2001).

Qualitative results of the effect of the number of dimensions rated have been promising (Ahmed et al., 1997; Lievens, 1998). Lievens (1998) concluded that the number of dimensions used should be small in order to increase convergent validity. Similar results have been found using quantitative methods to summarize data (Lievens & Conway, 2001; Woehr & Arthur, 2003). Lievens and Conway (2001) found that fewer dimensions led to significantly higher proportions of dimension versus exercise variance, and Woehr and Arthur (2003) found, in their meta-analysis, that convergent validity is higher for assessment centers measuring fewer dimensions. However, the authors of the latter study also found that divergent validity results were better for those assessment centers that used a larger number of dimensions. Gaugler and Thornton's (1989) lab study provides further evidence of this result. Assessors made ratings on three, six, or nine dimensions. Although those who rated nine dimensions demonstrated slightly less convergent validity than those who rated three, all three groups demonstrated high convergent validity overall. Discriminant validity results were not positive for any group, leading Gaugler and Thornton to suggest that a general factor is indeed underlying assessor ratings. However, this finding of higher convergent validity and lower discriminant validity for fewer dimensions is confounded by possible true interrelationships of assessment center dimensions, which were discussed above.

Although the effect of number of dimensions on the construct-related validity of assessment centers has resulted in only partial support (better convergent, but not necessarily divergent, validity), there is evidence that accuracy is influenced by number of dimensions rated. In Gaugler and Thornton's (1989) study, those assessors who rated

fewer dimensions classified behaviors more accurately and made more accurate ratings, leading the authors to conclude that rating fewer dimensions resulted in allowing the assessors to better handle the cognitive information processing demands of the assessment process. This result holds promise for research on the impact of number of dimensions on assessment center validity.

Cognitive and Information Processing Demands - Ability of Assessors

Although much of the research on the effect of information processing demands on the construct-related validity of assessment centers has focused on ways to improve the process to reduce these demands, some have questioned whether assessors, as humans, are even able to appropriately accomplish the task (Hoeft & Schuler, 2001; Jones, 1997; Kauffman et al., 1993; Lievens, 2001a, 2002; Turnage & Muchinsky, 1982). Is the lack of discriminant validity found in assessment centers partially due to assessors being unable to distinguish among dimensions?

Two studies have directly examined this question (Lievens, 2001b, 2002). Holding true performance consistent, Lievens (2001b) conducted a lab study that looked at assessor ability. Given a candidate who performs consistently across exercises for each dimension, can assessors give consistent across-exercise ratings? The results were positive; "when assessors rated candidates whose performances were designed to be relatively consistent across exercises, evidence of convergent validity was established" (p. 211). Further, to test the competing theories of whether the lack of divergent validity is due to assessors' inability to differentiate or to candidates' performances actually varying, assessors were presented with candidates whose performance fluctuated across

dimensions within exercises, and candidates whose performance did not. The results were again positive, with assessors able to find differences across dimensions for those candidates whose performances actually fluctuated.

Using generalizability theory analyses, Lievens found moderate levels of interrater reliability and a small variance component associated with the assessor main effect, indicating that assessors did not differ from one another - for example, it was not the case that some assessors were lenient while others were stringent. The results also indicated a low variance component for dimension (one dimension did not receive higher or lower ratings over other dimensions) and a moderate variance component for candidate (candidates differed somewhat from each other on the ratings). Given all these data, Lievens concluded that assessors' ratings are veridical and that assessors are capable of doing the task. In this controlled setting, assessors were able to demonstrate appropriate convergent and divergent validity. The results of the Lievens (2002) study echoed these conclusions: assessors are able to rate differentially and consistently when appropriate. In other words, they are able to accomplish the task of assessor.

Exercise and Dimension Design - Opportunity to Behave and Observe

When suggesting exercise and dimension design as possible reasons for the lack of construct-related validity found for some assessment centers, attention is usually focused on the opportunity, or lack of opportunity, to behave and observe. It has been proposed that an insufficient number of behaviors are being elicited by the exercises because there is not enough opportunity for the participants to demonstrate the behaviors or for the assessors to discover the behaviors (Ahmed et al., 1997; Brannick et al., 1989;

Bycio et al., 1987; Harris, Becker, & Smith, 1993; Highhouse & Harris, 1993; Jones, 1992; Joyce et al., 1994; Kleinmann & Koller, 1997; Reilly et al., 1990; Sackett & Dreher, 1982). Further, exercises may vary in their ability to elicit behaviors. For example, if an exercise only elicits a small number of dimension-relevant behaviors, then there is little information on which assessors can base their ratings. Therefore, the ratings of different dimensions within that same exercise are likely to be very similar. There is not enough information to distinguish between the different dimensions when there are a limited number of behaviors displayed. The result is high HTMM correlations, and thus a lack of construct-related validity evidence.

With respect to opportunity to observe, some dimensions are viewed as being harder to observe than other dimensions, specifically within each exercise. For example, there is ample opportunity to observe behaviors for the dimension organizing and planning within an in-basket exercise; however there is less opportunity to observe the same dimension within a leaderless group discussion. This discrepancy could lead to lower MTHM correlations, and thus a lack of construct-related validity evidence. This variability in opportunity to observe for dimensions across exercises also could lead to the halo effect sometimes found in assessor ratings (Hoeft & Schuler, 2001; Joyce et al., 1994; Kleinmann, Kuptsch, & Koller, 1996; Kudisch et al., 1997; Turnage & Muchinsky, 1982; Woehr & Arthur, 2003); which can in turn reduce the discriminant validity of assessment centers.

The lack of opportunity theory was directly examined by Kleinmann and Koller (1997) who had assessors rate whether the dimensions they were trained on were

observable in each of the exercises. Only the three highest rated dimensions on observability were included in the study. The confirmatory factor analysis results indicated a three-factor model as the best fit - with the factors matching the three dimensions rated. The large amount of variance in behaviors accounted for by the trait factors provided further evidence that the dimensions all made significant contributions. These results indicate that perhaps when assessment center dimensions are designed to be observable, construct-related validity of assessment centers is attainable.

More indirect support for this comes from a study that used generalizability theory to investigate the construct-related validity of an assessment center (Arthur et al., 2000). Within generalizability research, the variance accounted for by the Dimension by Exercise interaction indicates the extent to which dimension ratings vary by exercise. The authors concluded that the relatively low Dimension by Exercise contribution attained for this study meant that the dimensions were "generally assessable in all exercises" (p. 827). This finding is important in the context of the overall results that demonstrated both convergent and divergent validity of the assessment center. These studies show that perhaps a lack of opportunity to demonstrate or observe behavior may account for the lack of construct-related validity evidence often found for assessment centers in the extant literature.

Evaluation Method

Another design characteristic that has been posited as an explanation for the lack of construct-related validity is the evaluation method used in the assessment center. Evaluation method involves when the dimension ratings are completed, and although

there are some variations, there are two generally used evaluation methods: within-exercise and within-dimension (sometimes referred to as the AT&T method or the behavioral reporting method; Harris et al., 1993; Sackett & Dreher, 1982; Silverman et al., 1986; Thornton et al., 1997). The within-exercise process has assessors make dimension ratings for candidates immediately after each exercise. The within-dimension process, on the other hand, has assessors wait until completion of all exercises before ratings are made, and then the ratings are done by dimension across exercises. Because the within-exercise evaluation method may lead assessors to focus on the exercise and process information in terms of exercises, this method may lead to inflated HTMM correlations, and a lack of construct-related validity (Haaland & Christiansen, 2002; Joyce et al., 1994; Kauffman et al., 1993; Silverman et al., 1986). The within-dimension method instead focuses assessors on the appropriate factors, the dimensions, thus potentially increasing the construct-related validity.

Two applied studies evaluated the effect of evaluation method on the construct-related validity of assessment centers. Silverman et al. (1986) had assessors trained on and then rate candidates using either the within-dimension approach or the within-exercise approach. MTMM, ANOVA, and factor analysis results all provided evidence that the within-dimension method showed greater construct-related validity. Specifically, stronger convergent validity evidence (using MTMM) and higher discriminant validity evidence (using ANOVA) resulted for the within-dimension method, leading Silverman et al. to conclude that the variation in evaluation method led to observable differences in construct-related validity - perhaps because the different

methods "forced the raters to process and organize the assessment center data in different ways" (p. 573). However, finding fault with Silverman et al.'s study, Harris et al. (1993) also directly assessed the effect of evaluation method and found no differences in construct-related validity. Average MTHM correlations were almost identical for the different methods, and average HTMM correlations were higher than average MTHM correlations regardless of method. Confirmatory factor analyses echoed these results.

With conflicting applied study results, Robie, Osburn, Morris, Etchegaray, and Adams (2000) turned to a lab study to experimentally manipulate the evaluation process. In this study, each assessor made ratings either on one dimension across exercises (within-dimension method) or on one exercise for all dimensions (within-exercise method). The results provided strong support for the theory that within-exercise evaluation methods may be reducing the construct-related validity of assessment centers. MTHM correlations were higher than HTMM correlations for the within-dimension process, and the opposite was true for the within-exercise process. Further, confirmatory factor analyses showed a two-exercise factor solution for the within-exercise process and a four-dimension factor solution for the within-dimension process. Thus, the evaluation method may be a methodological artifact that influences construct-related validity results. In support of this conclusion, Woehr and Arthur's (2003) meta-analysis found that the construct-related validity of assessment centers is affected by rating approach, with the within-dimension method (or across-exercise, as the authors referred to it) yielding stronger construct-related validity evidence.

Evaluation method is just one more addition to the list of design and implementation methodological differences that may, or have been shown to, affect the construct-related validity of assessment centers. Another group of authors has turned to statistics to explain differences in construct-related validity results.

Statistical Approaches

A number of researchers have criticized the statistical approaches of some construct-related validity studies (Arthur et al., 2000; Donahue et al., 1997; Howard, 1997; Jones, 1992; Kleinmann & Koller, 1997; Lievens, 1998; Lievens & Keer, 2001; Sagie & Magnezy, 1997; Woehr & Arthur, 2003). Both MTMM and factor analysis results have come under attack, and some authors have questioned whether these methods are even appropriate ways to assess the construct-related validity of assessment centers.

Orthogonal rotations within exploratory factor analyses often have been used to model the factors of assessment centers with typical results reflecting exercise not dimension factors. However, the moderately sized intercorrelations of dimensions and the finding that it is unlikely that assessment center dimensions are independent of each other (Arthur et al., 2003) suggest that dimensions should not be estimated as orthogonal (Donahue et al., 1997; Woehr & Arthur, 2003).

Confirmatory factor analyses (CFA) also are not without problems in assessment center research. For example, Sagie and Magnezy (1997) suggested that there are potential problems with assuming non-zero correlations between exercises, and stressed the possibility that "the low AC construct validity tapped by this technique [confirmatory

factor analysis] resulted from an erroneous interpretation of method factors, and not from a lack of a genuine trait effect" (p. 103). Further, there often are estimation problems or inadmissible solutions that may lead to an underestimation of the construct-related validity of assessment centers (Kleinmann & Koller, 1997; Lievens & Keer, 1999, 2001).

From these problems, a CFA model that estimates correlated uniquenesses has been suggested (Binning, Adorno, & LeBreton, 1999; Kleinmann & Koller, 1997; Lievens & Keer, 2001; Sagie & Magnezy, 1997). The correlated uniqueness (CU) approach directly estimates only dimension factors - "exercise effects are inferred from the correlations among the error terms of ratings produced by the same exercise" (Lievens & Keer, 2001, p. 374). The results from using this method have been positive (Kleinmann & Koller, 1997; Lievens & Conway, 2000; Lievens & Keer, 1999, 2001; Sagie & Magnezy, 1997). Sagie and Magnezy's (1997) results showed that the expected five-dimension factor solution fit best, with evidence of convergent validity via factor loadings. Kleinmann and Koller (1997) applied the correlated uniqueness model to Bycio et al.'s (1987) data and found that the CU model fit best, with 35% of the variance accounted for by dimension factors. They then applied the method to a new assessment center with favorable construct-related validity results. Finally, in a review of the various statistical methods, Lievens and Conway (2000) compared different CFA models using within-exercise ratings of 24 assessment center studies. The different models compared were: correlated methods (CFA-CM; the traditional model that allows both dimensions and exercises to correlate); uncorrelated methods (CFA-UM; similar to the

traditional model, only exercises are estimated as uncorrelated); correlated uniqueness (CFA-CU; exercise factors are captured in the uniquenesses); and the direct product model (DP; which measures the interaction of traits and methods). Although all models produced, on average, adequate fit, the CFA-CU and DP models yielded a larger percentage of matrices with acceptable fit and proper estimation. The authors concluded that the parameter estimates of the CFA-CU model are most trustworthy, but cautioned that which model a researcher uses partially depends on whether the researcher believes exercises should be correlated.

Recently, however, researchers have shown that the CU model also is flawed. Lance, Lambert, Gewin, Lievens, and Conway (2004) re-analyzed the Lievens and Conway (2001) 34-study dataset and came to different conclusions. Specifically, it was suggested that use of the CU model upwardly biases the dimension effect to the extent that the exercises are not orthogonal (as the model must assume) and the exercise factor loadings are nonzero. The CU model omits these effects, leading to inflated estimates of convergent validity and underestimates of discriminant validity. Lance et al. investigated a one-dimension model that was not studied by Lievens and Conway. Using a quantitative approach to compare the one-dimension-correlated-exercises (1DCE) model to the one-dimension-correlated-uniqueness (1DCU) model, Lance et al. found evidence of upward bias for the CU model (higher dimension estimates). The authors concluded that, "compared with the 1DCE model, the 1DCU model overestimated dimension variance components by 93% and underestimated exercise variance components by 31% on the average" (p. 381). However, they further noted that

this conclusion assumes that the 1DCE model is in fact the correct model and that the 1DCU model is not - a conclusion that has not been proven.

Construct Misspecification

A third grouping of articles has focused on the theory that assessment centers demonstrate low construct-related validity because the dimensions within assessment centers are misspecified or incorrectly identified (Brannick et al., 1989; Chan, 1996; Donahue et al., 1997; Russell & Domm, 1995). In other words, assessment centers are measuring constructs other than those the designers of the assessment center intended.

Personality Factors

The simplest rationale of the construct misspecification theory is that the constructs being unintentionally measured in assessment centers are personality variables such as impression management or self-monitoring. Assessment centers are "working" (demonstrating criterion-related validity) because these personality factors are important constructs for good performance in both the assessment center and on the job. Therefore, candidates who are high on these personality factors will be rated high on assessment center dimensions and subsequently rated high in performance evaluations. The theory is that it is the personality factor, not the specified assessment center dimensions, that is driving the performance relationship. However, because these personality variables are not overtly being measured in the assessment center (they are not usually dimensions), assessment centers are demonstrating a lack of construct-related validity with respect to the intended dimensions. Simply put, "high criterion-related

validity implies that there must be construct validity in assessment centers but we have not yet identified the constructs" (Chan, 1996, p. 177).

Limited research on this issue has not found any support for this theory. For example, Arthur and Tubre (2002) directly assessed the relationship of self-monitoring, assessment center performance, and on-the-job performance. Self-monitoring in this study was operationalized as self-presentation. The rationale behind self-monitoring as a potential misspecified construct is that those who perform best in assessment centers do so because they are effective self-presenters. Further, assessment centers have high criterion-related validity because high self-monitors perform better both in the assessment center and on the job, but low construct-related validity is found for assessment centers because self-monitoring is not overtly measured/assessed. To test this, the authors had assessment center participants complete a self-monitoring inventory, and then the authors assessed the relationship between self-monitoring, assessment center performance, and performance on the job. The results showed a lack of support for the construct misspecification hypothesis - assessment center ratings and self-monitoring both were related to job performance, but self-monitoring was not related to assessment center ratings (OAR or individual dimension ratings). Although the results of this study are promising, a definitive answer cannot be determined based on one study alone; more research needs to be conducted.

Transparency of Dimensions

One suggestion for the lack of construct-related validity in assessment centers is that because dimensions are not revealed to participants (i.e., they are nontransparent),

the participants have to guess which dimensions are being rated and then act accordingly (Kleinmann, 1993; Kleinmann & Koller, 1997; Kleinmann et al., 1996; Lievens, 1998). This ability to judge the situation and then change one's behavior is a skill that is not overtly rated in most assessment centers; but it is covertly rated in that if a participant has this skill, then he/she should score better in the assessment center. The influence of this skill may be causing noise in the dimension ratings and therefore obscuring the convergent validity evidence for assessment centers. Perhaps by providing participants with information about, or making them aware of, the dimensions that will be rated, the impact of participants' ability to judge situations will be eliminated as a noise factor, and the ratings should demonstrate higher convergent validity. Kleinmann et al. (1996) found that when the dimensions were not given to the participants (i.e., nontransparency), the typical pattern of construct-related validity appeared - that is, the model that fit best had three correlated exercise factors and only one ability factor (oral communication). However, for the transparency condition, where participants were given the dimensions that would be rated, the model that fit best had three correlated exercise factors and three (only three dimensions were rated) correlated dimension factors; thus demonstrating construct-related validity.

With respect to the construct misspecification hypothesis, the theory is that perhaps people differ in their ability to determine which dimensions are being assessed and to alter their behavior accordingly. Indeed, there is evidence of this. Kleinmann (1993) had participants guess which dimensions were being rated and found that individuals varied in their ability to accurately identify which dimensions were being

measured. Thus, this ability to recognize and act could be a construct that is unintentionally being measured in assessment centers and, because it is not overtly measured, it could be contributing to the lack of construct-related validity evidence often found. Some researchers have therefore argued that dimensions and corresponding behaviors should be made transparent to participants of developmental assessment centers in order to increase the construct-related validity. However, Lievens and Conway's (2001) review of design characteristics that may increase the construct-related validity of assessment centers found no difference in construct-related validity evidence between those assessment centers that made dimensions transparent versus those that did not. Overall, the evidence with respect to transparency is limited and further research needs to be conducted before conclusions regarding transparency and the construct-related validity of assessment centers can be made.

Summary of Possible Reasons for Disparity of Construct-Related Validity Results

The preceding review of possible explanations for assessment center construct-related validity results suggests that the alleged lack of construct-related validity of assessment centers may be artifactual and not real. The explanations presented above are not an exhaustive list of all possible reasons why some studies demonstrate construct-related validity for assessment centers while others do not. Instead, they represent some of the factors that are relevant to the present study.

PURPOSE OF THIS DISSERTATION

Although much research has been conducted on the topic of the construct-related validity of assessment centers, a definitive conclusion has yet to be drawn. Each piece of future research on the topic, including the present study, should attempt to contribute evidence towards the resolution of this debate. The central question of this debate is whether the underlying factors of assessment centers are the dimensions rated. Although the above review of the literature suggests that the alleged lack of construct-related validity is indeed an artifact and not real, there are a number of researchers who have not come to the same conclusion. Therefore, there is still a need for studies to provide different ways of looking at the familiar problem. The present study attempts to provide new approaches to examining the old theories, as well as present some areas that have not been investigated.

Lack of Opportunity to Behave and Observe

Recall that one area suggested for the lack of construct-related validity in assessment centers involves the lack of opportunity to behave and observe. Only a limited amount of research, however, has investigated this possible explanation. The present study attempts to directly examine the impact of opportunity to behave and opportunity to observe on the construct-related validity of an assessment center.

Opportunity to Behave within Exercises

Opportunity to behave involves whether exercises provide enough opportunity for participants to display dimension-relevant behaviors. Opportunity to behave was approached from two perspectives: the potential level of opportunity to behave, and the

actual number of behaviors displayed. The potential level of opportunity to behave addressed whether the exercise provided participants with enough opportunity to demonstrate relevant behaviors, and was assessed using assessors' ratings. This was followed-up with using actual assessment center behavior counts. Thus, this study examined both opportunity to behave of, as well as the actual number of behaviors exhibited within, an exercise.

With respect to a lack of opportunity to behave, the theory is that exercises need to allow an adequate number of behaviors to be displayed in order to produce positive construct-related validity evidence, specifically with respect to the HTMM correlations. If only a small number of behaviors can be, or are, displayed, assessor ratings of different dimensions (within that exercise) could potentially be very similar because there is not much behavioral information on which to base the ratings. Therefore, it is predicted that:

> *Hypothesis 1a*: Exercises that are rated as having low ability to elicit dimension-related behaviors will demonstrate poor internal discriminant validity (using the MTMM framework) compared to exercises that are rated as having high ability to elicit behaviors.

> *Hypothesis 1b*: Exercises that enable the display of smaller numbers of behaviors will demonstrate poorer internal discriminant validity (using the MTMM framework) compared to exercises that enable the display of larger numbers of behaviors.

Hypothesis 1a addresses the impact of opportunity to behave, operationalized as assessors ratings of said opportunity, while Hypothesis 1b addresses the impact of the actual number of behaviors displayed.

Opportunity to Observe for Dimensions

Dimension ratings are susceptible to a lack of opportunity to observe. Opportunity to observe addresses the question of how observable the dimension is to assessors within the assessment center exercises. As with exercises, dimensions need to be defined such that an adequate number of behaviors can be observed within the exercises. Two characteristics for each dimension within each exercise were determined: the potential level of opportunity to observe, and the actual number of behaviors displayed. Similar to Hypothesis 1a, the potential level of opportunity to observe was determined by assessor ratings of the degree to which relevant dimension behaviors are able to be displayed within each exercise. In order for a dimension to provide positive construct-related validity evidence, there must be adequate opportunity to observe behaviors for *each* exercise. For example, if one is looking at the dimension team building, there must be enough behavioral information in each exercise in which team building is assessed (e.g., in-basket, leaderless group discussion) in order for internal convergent validity (assessed using MTMM) to be demonstrated. If the dimension (team building) does not demonstrate adequate opportunity to observe in one (or more) of the exercises (e.g., in-basket), the rating of that dimension for that exercise will be less accurate and therefore potentially different from the (team building) ratings of the other assessment exercises. This difference in ratings leads to lower convergent

validity, specifically with respect to MTHM correlations. The same line of reasoning applies for actual behavior counts. Consequently, it is hypothesized that:

> *Hypothesis 1c*: Dimensions that are evaluated as more observable for all relevant exercises will have higher internal convergent validity (using the MTMM framework) compared to those dimensions that do not allow for adequate opportunity to observe for all relevant exercises.

> *Hypothesis 1d*: If the number of displayed behaviors for a dimension is larger for all relevant exercises, internal convergent validity (using the MTMM framework) will be higher for that dimension, in comparison to those dimensions for which a smaller number of behaviors is consistently displayed and those dimensions for which the number of behaviors displayed is inconsistent across exercises.

Hypothesis 1c addresses the impact of opportunity to observe, operationalized as assessors ratings of said opportunity, while Hypothesis 1d addresses the impact of the actual number of behaviors displayed.

Evaluation Method

Another design characteristic that has been posited as an explanation for the lack of construct-related validity is the evaluation method used in the assessment center; within-dimension versus within-exercise. Although the applied research that has investigated this premise provided mixed results, a single lab study demonstrated some evidence that within-dimension evaluation methods result in better construct-related

validity. The present study attempts to provide more applied evidence by examining the effect of evaluation method on the construct-related validity of an operational assessment center.

The design of the assessment center used in this study provides a unique opportunity to compare the two evaluation methods within the same assessment center. For each participant there are two sets of ratings: those from an assessor who makes ratings immediately after the exercise is completed (within-exercise) and those from assessors who make dimension ratings across exercises once all the exercises are completed (within-dimension). This allows for a comparison of evaluation methods using the same candidate population in an applied setting. As the literature suggests, it is hypothesized that:

> *Hypothesis 2*: The within-dimension method will lead to more positive construct-related validity results versus the within-exercise evaluation method.

Note that the assessor ratings for the two evaluation methods are necessarily confounded in that the within-dimension assessors' ratings are based on the information that the within-exercise assessor provides. The broader implications of this are further addressed in the Discussion section of this paper.

Relationship of Exercises and Dimensions to Performance

One idea that has yet to receive much attention involves a comparative assessment of the criterion-related validity of dimension versus exercise scores (Arthur, Day, & Woehr, 2008; Connelly, Ones, Ramesh, & Goff, 2008; Jones & Klimoski, 2008).

Although evidence of criterion-related validity does not directly indicate evidence of construct-related validity, a comparison of the relationships of exercises and dimensions to performance may shed some light on the debate. Specifically, if, as many researchers suggest, assessment center ratings represent exercises and not dimensions then one would expect exercise/performance correlations to be higher than dimension/performance correlations. However, if assessment center ratings reflect dimensions as they are designed to do, then the opposite should hold true.

Given the so-called exercise effect, many researchers also have called for a shift from using dimension scores to using exercise scores to assess "roles" (Hoeft & Schuler, 2001; Klimoski & Brickner, 1987; Lance, 2008b; Lance et al., 2000; Robertson et al., 1987; Russell & Domm, 1995; Sackett & Dreher, 1982; Sackett & Harris, 1988; Silverman et al., 1986). In this scenario, exercise scores could be used for both feedback and prediction of performance within a role congruency context. This "shift" requires an examination of the criterion-related validity of exercise scores. Two studies have taken initial steps towards looking at this issue (Lance et al., 2000; Lance, Foster, Gentry, & Thoresen, 2004). Both studies examined the situational specificity hypothesis that states that exercise effects do not represent method or measurement biases, but are instead situationally specific performance factors. Within this context, the authors presented data that showed significant positive relationships between exercise factor scores and external measures (e.g., job knowledge, reading comprehension), and most importantly for this study, supervisor ratings of job performance (Lance et al., 2004). Although these studies provide some evidence of a relationship between exercises and

performance, there are two main differences between these studies and the present one. First, these studies do not investigate the exercise/performance relationship relative to the dimension/performance relationship. Second, the present study examines the relationship at the rating level as opposed to the factor level.

Recall that a recent meta-analysis (Arthur et al., 2003) found that individual dimension ratings demonstrated validity coefficients equal to or higher than the .37 found by Gaugler et al. (1987) for OAR; this highlights the need of assessment center research to move away from solely examining the OAR, and to increase examination of dimension- and exercise-level information. The present study does just that by comparing the predictive validity of individual exercise scores to individual dimension scores. If the true factor structure of the assessment center is comprised of exercise factors, then individual exercise scores should be better predictors of performance than dimension scores. If, on the other hand, dimensions truly are the constructs underlying assessment ratings, then the opposite should hold.

The design and implementation of the assessment center used in this study is based on research findings regarding construct-related validity evidence. Specifically, psychologists were used as assessors and FOR training was used to prepare assessors for the task. Also, an attempt was made to reduce the cognitive demands on assessors by creating distinct, behaviorally-defined dimension definitions that remained consistent across exercises, and the number of dimensions was kept to a manageable number. These steps were taken in order to reduce the potential impact of design characteristics

on the construct-related validity of the assessment center, and provide a clearer picture of the evidence.

Taking the design characteristics of this assessment center and the research to date on the construct-related validity of assessment centers into account, it is hypothesized that:

> *Hypothesis 3*: The dimension/performance relationships will be stronger than the exercise/performance relationships.

Stability of Dimensions across Exercises

A relatively new idea in the study of the construct-related validity of assessment centers has been to look to the personality literature (cf. Haaland & Christiansen, 2002; Tett, 1999; Tett & Schleicher, 2001), specifically the trait versus state distinction, to explain and assess the general mixed findings. Although fully examining this new approach is beyond the scope of the present research, an exploratory investigation of some basic concepts of personality may provide more evidence for the debate. This exploratory investigation looks at the differences in stability of dimension ratings across exercises.

The use of factor analysis and internal MTMM matrices indicates that many researchers have been treating, if not viewing, assessment center dimensions as personality traits (Hoeft & Schuler, 2001; Howard, 2008; Lance, 2008a). The extent to which an assessment center dimension approaches a personality trait should affect internal construct-related validity, in that the more a dimension is like a trait, the more stable the ratings are expected to be. In other words, those dimensions that are defined

as more similar to personality traits and that represent clusters of behaviors that are not bound by situational constraints would be more likely to demonstrate the cross-situational consistency wanted in MTMM analyses. Those dimensions that are more situationally based, more dependent upon the situation/exercise, would be less likely to demonstrate cross-situational consistency and thus construct-related validity. Personality literature indicates that the situation, or in the case of assessment centers, the exercise, can constrain behavior. The question for assessment center research is whether those dimensions that can transcend the constraints, those that are more like traits, are more likely to yield construct-related validity versus those that are bound by the situational constraints (i.e., exercises).

An examination of dimension definitions (within the assessment center used in this study) reveals that dimensions may differ with respect to their stability across exercises. Some dimensions are general and should be consistent across exercises, whereas other dimensions appear to be more situationally specific. For example, oral communication has been found to demonstrate higher consistency versus specificity (Hoeft & Schuler, 2001; see also Howard, 2008). In a study of a developmental assessment center, Engelbrecht and Fischer (1995) found that synthesis and judgment did not change as a result of the assessment process. The authors concluded that perhaps these two dimensions, because they are cognitive in nature, are more similar to enduring traits. Other authors also have suggested that dimensions such as decision-making and problem-solving, again being cognitive in nature, are harder to change and are more stable (Boehm, 1985; Connelly et al., 2008; Turnage & Muchinsky, 1982).

Building off of this research and considering the dimension definitions within the present assessment center, problem solving and oral communication would appear to be more similar to stable traits and therefore demonstrate more consistency. Two of the other three dimensions, organizing and planning and team building, are, on the other hand, less cognitive in nature and have facets that can be learned or changed. For example, there are specific behavioral steps that a person can take to improve his/her organizing and planning skills - such as, keeping time during group sessions, stacking letters and memos based on priority, using the calendar when making meetings, etc. These behaviors can be learned and improved upon more easily than those under oral communication and problem solving.

Further, while problem solving and oral communication are more consistent, the other two dimensions have facets that are seemingly more dependent on context. The behaviors displayed for these two dimensions may have been previously learned only within one context and not another. So, for example, an individual could demonstrate strong team building in the leaderless group discussion but not in the in-basket task because he/she has only learned about the benefits of team building within a group setting and has not had experience with team building in a written context, such as letters, memos, and the like. This person, demonstrating strong team building within the group exercises but low team building otherwise, would receive inconsistent scores across exercises. These two dimensions (organizing and planning, and team building), within the context of this assessment center, appear to be more situationally specific and therefore less stable.

It is expected that the more stable dimensions (oral communication and problem solving), those that are more similar to traits, will have higher MTHM correlations, indicating convergent validity. The more situationally specific dimensions (team building, and organizing and planning) are expected to demonstrate lower MTHM correlations, and thus, lower internal construct-related validity. This comparison of stable versus situationally specific dimensions is similar to the trait versus state distinction made in personality research.

Summary of Dissertation Objectives

The four main objectives of this dissertation address the construct-related validity of an assessment center from different angles. The first hypothesis investigates whether the opportunity to behave and the opportunity to observe, as rated by assessors and assessed via a counting of behaviors, influence discriminant and convergent validity, respectively. The second hypothesis addresses the debate over evaluation method and examines which method, within-exercise or within-dimension, yields more favorable internal construct-related validity evidence. The third hypothesis explores the call for exercise scoring in assessment centers and compares the criterion-related validity of exercise versus dimension scores within the same assessment center. Finally, the fourth objective looks at the relationship of the stability of the dimension and the internal construct-related validity, specifically convergent validity evidence.

METHOD

Participants

Data for this study were obtained from seven administrations of an evolving developmental assessment center, spanning five years. Two administrations were conducted for low to mid-level managers in a state agency (N = 31). The remaining five administrations were conducted in a professional graduate school of public administration that prepares students for leadership and managerial positions in government and public service (N = 108). The seven administrations yielded a total sample size of 139 participants.

Materials

The Assessment Center

The assessment center was originally developed for the state agency and was then modified for administration in the professional graduate school of public administration. It was developed using a content-related validation strategy that included job analyses, identification of work behaviors and knowledge, skills, abilities, and other characteristics (KSAOs), identification of behavioral dimensions related to the KSAOs and work behaviors, and finally, development of exercises to tap the specified behavioral dimensions. The resulting assessment center (Arthur, 1997; Arthur, 2001) was designed to measure five behavioral dimensions - oral communication, influencing others, team building, problem solving, and organizing and planning - using three exercises. The three exercises were: (1) a competitive resource allocation exercise (leaderless group discussion); (2) an in-basket exercise **followed by an oral interview** to

answer questions concerning the in-basket; and (3) a non-competitive management problem exercise (a second leaderless group discussion). Dimension definitions are provided in Appendix A.

Assessors and Their Training

Across the seven administrations, 35 different assessors (17 males and 18 females) were used. Assessors included persons who had earned their doctorate in industrial/organizational (I/O) psychology and advanced level I/O or social psychology graduate students. The participant-to-assessor ratio was either 1:1 or 2:1.

A two-day FOR training program was used to prepare assessors for the assessment center. Prior to training, assessors received and reviewed a training manual. The manual and training session provided assessors with information regarding the assessment center, dimensions, and exercises. The first step in the training was to familiarize the assessors with what an assessment center is and does. After a general overview, the specific process used in the current assessment center was outlined, and the exercises used were explained. Examples of the exercises were included in the training manual. Next, the assessment center dimensions were described and discussed in detail. The descriptions included behaviors that assessors may observe in the assessment center, as well as behavioral anchors for each dimension rating scale (see Appendix B). For each dimension, the same behavioral anchors were used regardless of the exercise. General information for observing behaviors was provided to assessors in order to emphasize the use of objective, verifiable observations in note taking and classification.

Once assessors processed and understood the exercises, dimensions, and rating scales, they participated in mock exercises. First, assessors watched a videotape of the competitive leaderless group discussion (LGD). During the LGD, assessors took notes on the same two participants. At the completion of the LGD, assessors categorized their written behavioral observations into the appropriate dimensions. Assessors then made ratings on the relevant dimensions using the dimension definitions and rating scales and the classified behaviors. Assessors were aware that they should be able to provide specific behavioral evidence for or justification of their ratings if needed. Once all individual ratings had been made, the ratings were shared with the group and discussed until consensus was reached. This discussion allowed for assessors to develop a common frame of reference. This process occurred similarly for the in-basket exercise. Assessors received a completed in-basket exercise and were to record behaviors found in the in-basket. For this exercise, assessors also were able to practice the interview portion, allowing them the opportunity to clarify any issues in the participants' written responses. Assessors then classified the behaviors and provided ratings for the relevant dimensions. Again, ratings and behaviors were presented and discussed in order to define a common frame of reference for the assessors.

The next segment of training was an overview of the overall rating process across the assessment center. This included a review of and tips for conducting the consensus meeting and reaching consensus. It also included a discussion of the evaluation method. For each exercise, there would be one primary assessor and one or more secondary assessors. The primary assessors would be using a within-exercise evaluation approach

(i.e., they would make their ratings after each exercise). The secondary assessors would be using a within-dimension evaluation approach. This occurred during the consensus meetings where assessors were to work across exercise, completing discussions and ratings for each dimension before moving on to the next one. Finally, because this was a developmental assessment center, feedback was an important aspect of the assessor's role. During training, methods of providing appropriate and relevant feedback were provided, and assessors had the opportunity to practice a feedback session. Assessors also received examples of feedback summaries to assist with the written portion of the feedback.

Performance Data

Performance data were collected as part of the overall assessment center process. These data were collected between six and twelve months after completion of the assessment center. For the state agency, performance ratings were collected from self, supervisors, peers, and direct subordinates. For the professional graduate school of public administration, ratings were collected from self, internship supervisors, professors from whom the participants had taken classes, and students in the participants' incoming class (i.e., peers). Ratings were anonymous, with the exception of self-ratings.

The performance measure was developed to elicit perceptions of the candidate's behaviors on leadership dimensions that directly mapped onto the assessment center dimensions (i.e., oral communication, influencing others, team building, problem solving, and organizing and planning). An Overall Effectiveness rating was collected for each dimension. Persons were asked to rate the candidate's effectiveness/success on

each dimension using a five-level scale that ranged from "very successful/effective" to "very unsuccessful/ineffective." There also was a place to indicate if the rater felt that he/she had insufficient information to make a rating.

Assessor Measure of Opportunity to Behave and Observe

In order to test Hypotheses 1a and 1c, a web-based measure was sent out to assessors via email requesting ratings of the opportunity to observe provided by each dimension/exercise combination (see Appendix C). The measure was sent out between six and ten years after assessors participated in the assessment center. Of the 34 total assessors, 31 completed the measure, yielding a 91% return rate. The measure asked assessors to rate, on a scale of 1 to 7, the level of opportunity to observe for each dimension/exercise combination. Assessors also provided summary ratings for each exercise – the overall level of opportunity to observe for each exercise across dimensions. The assessors' responses supplied the data necessary to determine the relative levels of potential opportunity to behave and observe for exercises and dimensions, respectively.

Behavior Counts for Dimension/Exercise Combinations

Three research assistants, who received independent study course credit, counted the number of behaviors listed on assessor reports for each dimension/exercise combination. The assistants were trained as a group and received detailed information on the dimension definitions and exercise descriptions, as well as information on the assessment center as a whole. Once an understanding of the assessment center was reached, the assistants were trained to recognize and identify what constitutes an

appropriate behavior, how behaviors should be counted, and how to distinguish between the different dimensions. Group discussion over numerous examples provided the assistants with the same frame of reference. They then counted behaviors for some participants on their own and the results were discussed as a group and discrepancies reconciled.

Each assistant counted the behaviors for every participant within three of the four assigned years. (Behavior-level data was not available for the state agency administrations or the first professional graduate school of public administration administration.) This yielded at least two independent counts per participant per year, and allowed for verification. Due to academic semester circumstances, the assistants were not able to convene and discuss any disagreements for reconciliation. Therefore, I reviewed any discrepancies, provided my own count of the behaviors, and finalized the data appropriately. Overall, 50% of the counts needed to be reconciled for years 1998 and 1999. For year 2000, all three assistants counted the behaviors, therefore reconciliations were needed for only 18% of the data. Unfortunately for 2001, it appeared that one of the two assistants assigned to the year did not appropriately perform the counting task (e.g., there were clear errors where it appeared the assistant merely copied what the other assistant had counted). For that year, I counted the behaviors for all participants, yielding a 36% discrepancy rate. For all data points, at least two persons agreed on the number of behaviors displayed.

Procedure

Each assessment center candidate participated in each of the three exercises and was evaluated on the five dimensions discussed above. Assessors were divided into groups of three or four, with each group assigned to observe and evaluate a group of four to six participants. For the two LGD exercises this included sitting towards the back of the room, away from the participants, so as not to be obtrusive to the process. Although towards the back of the room, assessors were situated such that they could readily observe the behaviors of their assigned participants. Assessors each observed and recorded behavior for one or two assessment center participants. This involved the assessor chronologically recording behaviors observed during the planning periods, presentations, and group discussion. Assessors were trained to record only that which was observable, and to be as detailed and descriptive as possible in their recordings.

Each assessor was *assigned* to observe different participants for each exercise (primary assessor); however, all assessors in the group were present during each LGD. Upon completion of each LGD, primary assessors categorized the recorded behaviors into relevant dimensions and made ratings for the appropriate dimensions. For the in-basket exercise, assessors reviewed and evaluated the in-basket items of one or two participants and then conducted the in-basket interview individually (i.e., no other assessors observed the interview process). As with the group exercises, primary assessors (the assessor who conducted the interview) recorded and categorized behaviors elicited from both the in-basket interview and the in-basket items themselves, and made

ratings on the relevant dimensions. Appendix D provides the linkages between each dimension and exercise in the assessment center.

Once all exercises were completed, assessors met within their assessor groups for consensus meetings. Within a participant, each dimension was discussed across exercises. Further, each participant was evaluated individually and completely before evaluation of another participant could begin. The consensus process for an individual participant began with the appropriate primary assessor reading the classified behaviors for the relevant dimension (e.g., team building) for the appropriate exercise (e.g., in-basket). All other assessors made ratings for that dimension (team building) on that exercise (in-basket) based on this verbatim listing of observed behaviors, as well as the assessor's own observations. These initial ratings were made independently and without discussion. This method was repeated for all exercises relevant to the dimension in question (team building). Next, assessors each made an independent dimension-level rating based on all the information heard and observed. These dimension scores were presented to the group and discussed until consensus was reached. The final dimension score was the score reached at consensus. This process was repeated for the remaining four dimensions for the selected participant. The entire sequence was then applied to the rest of the assessment center participants assigned to the group of assessors.

RESULTS

Hypothesis 1: Lack of Opportunity to Behave and Observe

Hypothesis 1 involved the investigation of the opportunity to behave and observe for exercises and dimensions, respectively. Opportunity to behave and opportunity to observe both were assessed two different ways: via an assessor measure of opportunity to observe and counting the number of behaviors per dimension/exercise combination. Both approaches were described in the Method section.

Hypothesis 1a: Opportunity to Behave within Exercises – Assessor Ratings

Hypotheses 1a stated that exercises that are rated as having lower ability to elicit dimension-related behaviors will demonstrate poorer internal discriminant validity (using the MTMM framework) compared to exercises that are rated as having higher ability to elicit behaviors. For this hypothesis, the relative levels of opportunity to behave within exercises were determined by responses from the assessor measure of opportunity to behave. Table 1 provides the results of both the average exercise ratings across dimensions and the overall exercise summary ratings. (Recall that assessors were asked to rate each exercise on its overall ability to elicit behaviors across all dimensions – these ratings provided the data for "Average summary rating" in the table.)

Looking at the average rating across dimensions, the policy analysis exercise was rated as having the highest ability to elicit behaviors followed by the resource allocation exercise, although the difference between these two is small. The in-basket was rated as having the lowest ability to elicit behaviors. The summary ratings yielded similar results with the in-basket exercise receiving the lowest ratings of opportunity to behave.

However, in contrast to the ratings across dimensions, the resource allocation exercise was rated as having a relatively higher level of opportunity to behave than the policy analysis exercise, although again the difference is small.

Table 1
Assessors' Ratings of Opportunity to Behave within Exercises

	Resource Allocation	In-basket (with interview)	Policy Analysis
Average rating across dimensions	5.68	4.61	5.85
Average summary rating	6.23	4.77	5.97

Note. $N = 31$. Rating scale ranged from one to seven.

Support for Hypothesis 1a would be demonstrated by relatively lower HTMM correlations the higher the exercise is rated, and thus, relatively higher HTMM correlations the lower the exercise is rated. Functionally, the in-basket exercise, having the lowest ratings, should have the highest HTMM correlations. The results, presented in Table 2, generally showed the reverse pattern; the two LGDs demonstrated higher HTMM correlations than the in-basket exercise for all dimension pairings except problem solving with organizing and planning for the policy analysis exercise (.50 for in-basket versus .42 for policy analysis). Averaging across all dimension pairings produced the same result (.51 and .48 for the LGDs vs. .38 for the in-basket).

Table 2
Heterotrait-Monomethod Correlations for Exercises

Dimension Pairings	Resource Allocation	In-basket (with interview)	Policy Analysis	Two LGDs Combined
Oral Communication – Influencing Others	0.56		0.55	0.55
Oral Communication – Team Building	0.48	0.39	0.57	0.53
Oral Communication – Problem Solving	0.45	0.29	0.36	0.41
Oral Communication – Organizing and Planning	0.49	0.24	0.44	0.47
Influencing Others – Team Building	0.53		0.48	0.51
Influencing Others – Problem Solving	0.56		0.45	0.51
Influencing Others – Organizing and Planning	0.52		0.53	0.53
Team Building – Problem Solving	0.42	0.41	0.44	0.43
Team Building – Organizing and Planning	0.52	0.47	0.50	0.51
Problem Solving – Organizing and Planning	0.52	0.50	0.42	0.47
Average[a]	0.51	0.38	0.48	0.49

Note. k = 427 for Resource Allocation, 432 for In-basket, and 429 for Policy Analysis, where k is the number of data points. Influencing others was not assessed in the in-basket exercise.
[a] Indicates average correlation across all dimension pairings.

Analysis of the two LGDs necessarily produced mixed results because the two overall ratings, the average rating across exercises and the exercise summary rating, did not yield the same ranking for the two LGDs. This, combined with the weak contrast between the two LGDs on both overall ratings (see Table 1), did not lead to a clear pattern of results in examining the HTMM correlations. Therefore, to obtain a more

parsimonious interpretation, the two LGDs were combined for further analysis of this hypothesis.

Using this approach, support for Hypothesis 1a would be demonstrated by relatively low HTMM correlations for the combined LGD group, and correspondingly higher HTMM correlations for the in-basket exercise. The results, presented in Table 2, showed the reverse pattern; combined, the LGDs demonstrated higher HTMM correlations than the in-basket exercise. This held for all dimension pairings except problem solving with organizing and planning, where the combined LGD group had a slightly lower correlation than the in-basket exercise (.47 vs. .50). Averaging across all dimension pairings produced the same result (.49 for combined LGD vs. .38 for in-basket exercise).

Although the hypothesis is stated in relative terms, the data also were analyzed using more absolute benchmarks. The two benchmarks used were Hemphill's (2003) empirical guidelines for effect sizes, and three assessment center construct-related validity articles that provided summary statistics for MTMM matrices (Bowler & Woehr, 2006; Jones, 1992; Lievens, Chasteen, Day, & Christiansen, 2006). Hemphill (2003) suggested three levels of effect sizes for correlation coefficients within psychological studies: < .20 is a small effect, .20 to .30 is a medium effect, and > .30 is a large effect. Using this benchmark, the in-basket exercise did demonstrate "large" or "medium" effect sizes, which is consistent with the hypothesis; however, neither LGD demonstrated the "small" effect sizes that were hypothesized.

Three assessment center articles provided summary statistics for assessment center MTMM correlations that can be compared to the present study's data. Bowler and Woehr (2006) combined MTMM matrices of past assessment center studies into one large MTMM matrix and analyzed it, providing MTHM, HTMM, and HTHM summary statistics. These values are listed in Table 3. In a relatively qualitative approach, Jones (1992) provided average correlations, corrected for sample size, using data from 10 MTMM assessment center studies. These values also are presented in Table 3. Lievens et al. (2006) aggregated correlations across 30 MTMM matrices to provide average HTMM and MTHM correlations by exercise and dimension respectively. The authors reported the statistics based on trait activation theory, so the summary data for exercises was broken into two groupings, dissimilar and similar. According to the article, "results for similar links were derived from ratings between two dimensions that shared a link to the same personality trait, whereas results for dissimilar links involved two dimensions that did not share a link to any personality trait" (p. 253). Both values are provided in Table 3. (For Hypothesis 1a the focus is on HTMM correlations, so the middle column, labeled HTMM, is the relevant one.)

Using these three studies as benchmarks, the combined LGD group did, in general, yield lower HTMM correlations than the benchmark studies found (one notable exception being Lievens et al.'s (2006) Dissimilar LGD Competitive value of .44 for which only two HTMM correlations of the present study were found to be lower). However, the in-basket exercise demonstrated lower HTMM correlations as well.

Table 3
Benchmark MTMM Summary Statistics Based on Assessment Center Construct-Related Validity Articles

Article/Source	MTHM	HTMM	HTHM
Bowler & Woehr (2006)	0.25	0.53	0.20
Jones (1992)	0.39	0.58	0.25
Lievens, Chasteen, Day, & Christiansen (2006)			
– Dissimilar Links			
LGD – Competitive		0.44	
LGD – Cooperative		0.58	
In-basket Exercise		0.63	
– Similar Links			
LGD – Competitive		0.58	
LGD – Cooperative		0.58	
In-basket Exercise		0.62	
Lievens, Chasteen, Day, & Christiansen (2006)			
– Low Trait Activation Potential			
Problem solving (Openness)	0.29		
Team building (Agreeableness)	0.27		
Organizing and Planning (Conscientiousness)	0.22		
– High Trait Activation Potential			
Problem solving (Openness)	0.33		
Team building (Agreeableness)	0.30		
Organizing and Planning (Conscientiousness)	0.31		

In summary, both the relative and absolute analyses produced results that did not support Hypothesis 1a. The exercises that were rated as providing a higher opportunity to behave should have produced lower HTMM correlations (discriminant validity) compared to the exercises rated as having low opportunity to behave. This pattern did not hold for the present study.

Hypothesis 1b: Behavior Counts within Exercises

Hypothesis 1b stated that exercises that enable the display of smaller numbers of behaviors will demonstrate poorer internal discriminant validity (using the MTMM framework) compared to exercises that enable the display of larger numbers of behaviors. The average behavior counts of the research assistants (described in the Method section) were used for this hypothesis. The average number of behaviors displayed per dimension was 8.38 for resource allocation, 11.05 for the in-basket, and 7.26 for policy analysis. Relatively speaking, the in-basket exercise had the highest number of behaviors listed per dimension, followed by the resource allocation exercise, and then the policy analysis exercise (although the contrast between the two LGDs is relatively small in magnitude). Note that the rank ordering here is somewhat at odds with the ordering of exercises found in Hypothesis 1a where the in-basket exercise was rated as the lowest. This issue is discussed in a later section of this paper.

As with Hypothesis 1a, support for Hypothesis 1b would be established by lower HTMM correlations for exercises with a higher number of behaviors demonstrated, and higher HTMM correlations for those with lower number of behaviors displayed. Functionally, the desired outcome is for the in-basket exercise to have the lowest HTMM correlations, followed by the resource allocation exercise, and then the policy analysis exercise. This pattern partially held for the present data (Table 4); both LGDs produced higher HTMM correlations than did the in-basket exercise for all dimension pairings except problem solving with organizing and planning, where the in-basket had a

higher correlation than the policy analysis exercise (.50 vs. .42). Averaging across all dimension pairings produced the same result.

Table 4
Heterotrait-Monomethod Correlations for Exercises Based on Average Number of Behaviors per Dimension Displayed

Dimension Pairings	**High** In-basket (with interview)	Resource Allocation	**Low** Policy Analysis	Two LGDs Combined
Oral Communication – Influencing Others		0.56	0.55	0.55
Oral Communication – Team Building[a]	0.39	0.48	0.57	0.53
Oral Communication – Problem Solving[a]	0.29	0.45	0.36	0.41
Oral Communication – Organizing and Planning[a]	0.24	0.49	0.44	0.47
Influencing Others – Team Building		0.53	0.48	0.51
Influencing Others – Problem Solving		0.56	0.45	0.51
Influencing Others – Organizing and Planning		0.52	0.53	0.53
Team Building – Problem Solving	0.41	0.42	0.44	0.43
Team Building – Organizing and Planning	0.47	0.52	0.50	0.51
Problem Solving – Organizing and Planning	0.50	0.52	0.42	0.47
Average[b]	0.38	0.51	0.48	0.49

Note. k = 432 for In-basket, 427 for Resource Allocation, and 429 for Policy Analysis, where k is the number of data points.
[a] Indicates dimension pairings that yielded significantly different correlations for the in-basket versus combined LGD group. [b] Indicates average correlation across all dimension pairings.

The pattern of results in comparing the two LGDs to each other was not as clear. However, given that the average behavior count difference between them was not as

stark as the difference between the LGDs and the in-basket exercise (8.38 for resource allocation, 11.05 for the in-basket, and 7.26 for policy analysis), one would not expect a large contrast in correlations to emerge. Therefore, in order to obtain a more parsimonious result, the two LGDs were combined for the additional follow-up analyses.

When comparing the combined LGD group to the in-basket, the pattern of correlations was in the hypothesized direction (see Table 4); therefore the data were analyzed further using Steiger's (1980) formula for testing the equality of two dependent correlations with no index in common. Statistically significant differences ($p < .05$) between the High (in-basket) and combined LGD groups were found for three dimension pairings: oral communication with team building, oral communication with problem solving, and oral communication with organizing and planning. Using Fisher Z transformations and confidence intervals (Myers & Well, 1991) a statistically significant difference also was found for the overall average ($z = 1.95$, $\alpha < .05$; lower bound of confidence interval around in-basket average correlation = .30 and upper bound = .46). These findings demonstrate support for Hypothesis 1b, as the exercise with the most observed behaviors per dimension (the in-basket exercise) yielded statistically significant lower HTMM correlations (i.e., showed better discriminant validity) than the exercises with a lower number of observed behaviors per dimension (the two LGDs combined).

Although the hypothesis is stated in relative terms, the data also were analyzed using more absolute benchmarks. As with Hypothesis 1a, the two benchmarks used were Hemphill's (2003) guidelines for effect sizes, and three assessment center

construct-related validity articles that provided summary statistics for MTMM matrices (Bowler & Woehr, 2006; Jones, 1992; Lievens, Chasteen, Day, & Christiansen, 2006). Using Hemphill's benchmarks, the combined LGD grouping did demonstrate "large" effect sizes, which is consistent with the hypothesis, but the High grouping did *not* have the "small" effect sizes that were hypothesized. However, two of the six correlations (oral communication with problem solving and oral communication with organizing and planning) did fall into the "medium" effect size category providing partial support for the hypothesis.

Table 3 provides the summary statistics for the three assessment center MTMM benchmark studies. Using these three studies, the High grouping for Hypothesis 1b did produce lower HTMM correlations for all dimension pairings, as well as for the average correlation, which is consistent with the hypothesis. Further, although overall the combined LGD grouping demonstrated lower HTMM correlations as well, there were exceptions. One correlation was higher than the Bowler and Woehr (2006) standard, albeit only slightly (.55 for the oral communication with influencing others pairing versus .53). Nine correlations, including the overall average, were higher than Lievens et al.'s (2006) dissimilar link competitive LGD correlation of .44. This provided further support for Hypothesis 1b.

Taken together, the relative and absolute analyses produced results that generally supported Hypothesis 1b. The exercise that had a higher average number of behaviors per dimension (and thus, a higher opportunity to behave) produced lower HTMM

correlations (discriminant validity) compared to the exercises that had lower average number of behaviors per dimension (and thus, a lower opportunity to behave).

Hypotheses 1c and 1d: Opportunity to Observe for Dimensions

Hypotheses 1c and 1d involved the level of opportunity to observe for dimensions. These hypotheses stated that dimensions that are either evaluated as more observable or yield more behaviors for all relevant exercises will have higher internal convergent validity (using the MTMM framework) compared to those dimensions that do not allow for adequate opportunity to observe for all relevant exercises. Along with the focus on dimensions instead of exercises, these hypotheses differ from Hypotheses 1a and 1b in that they call for dimensions that are *consistently* high across all relevant exercises, as opposed to high on average. Therefore, two groupings were needed: dimensions that were rated as high (or yielded high counts of behaviors) for all exercises versus those dimensions that were either rated as low (or marginal) for all dimensions or were inconsistently rated (e.g., high for some exercises and low for others).

Hypothesis 1c

To analyze Hypothesis 1c, the relative levels of opportunity to observe for dimensions were determined using the responses from the assessor measure of opportunity to observe (see Table 5). These data produced two groupings of dimensions: a “High” opportunity to observe group and a “Low” opportunity to observe group.

Table 5
Assessors' Ratings of Opportunity to Observe for Dimensions

	Oral Communication	Influencing Others	Team Building	Problem Solving	Organizing & Planning
Resource Allocation	6.61	6.74	4.16	5.29	5.61
In-basket (with interview)	4.26		6.71	6.29	3.19
Policy Analysis	6.52	6.29	4.35	5.94	6.13
Mean rating across exercises	5.80	6.52	5.08	5.84	4.98

Note. $N = 31$. Rating scale ranged from one to seven. Influencing others was not assessed for the in-basket exercise, therefore there is no rating provided for that combination.

The High group included influencing others and problem solving because these dimensions were, relatively speaking, consistently high across all relevant exercises. The Low group included the other three dimensions: oral communication, team building, and organizing and planning. All three of the dimensions in the Low group were placed there because of inconsistent ratings; that is, they were rated relatively high on some exercises and relatively low on others. Relatively speaking, influencing others and problem solving had consistently higher rated opportunity to observe versus the other three dimensions.

Table 6
Monotrait-Heteromethod Correlations for High and Low Groupings Based on Assessors' Ratings of Opportunity to Observe for Dimensions

Exercise Pairings	**High Group** Influencing Others and Problem Solving	**Low Group** Oral Communication, Team Building, and Organizing & Planning
Resource Allocation – In-basket (with interview)	0.34	0.44
Resource Allocation – Policy Analysis	0.58	0.62
In-basket (with interview) – Policy Analysis	0.41	0.49
Average[a]	0.44	0.52

Note. k = 424 to 429 for both groups, where k is the number of data points.
[a] Indicates average correlation across all exercise pairings.

Support for Hypothesis 1c would be demonstrated by relatively high MTHM correlations for the High group, and lower MTHM correlations for the Low group. The results, shown in Table 6, indicated a reverse pattern; combined, influencing others and problem solving showed lower MTHM correlations than oral communication, team building, and organizing and planning combined. This held for all exercise pairings, as well as for the overall average across all exercise pairings.

Although the hypothesis was stated in relative terms, the data were analyzed further using the same absolute analyses as conducted previously in Hypotheses 1a and 1b. Using Hemphill's benchmark, the High group did demonstrate "large" effect sizes, which was consistent with the hypothesis; however, the Low group also demonstrated "large" effect sizes, which did not fit the hypothesis.

Turning to the summary statistics data, Lievens et al. (2006) reported their findings based on trait activation theory, so the summary data for dimensions was broken into two groupings, low trait activation potential of exercises and high trait activation potential of exercises. The authors noted that "results for high trait activation potential were derived from ratings between two exercises both high in activation potential for the same trait" (p. 253), while "results for low involved at least one exercise that was not high in trait activation potential for that trait" (p. 253). Further, the assessment center dimensions were clustered into the Big Five personality traits. Organizing and planning was specifically listed under Conscientiousness, and problem solving was specifically listed under Openness. Team building was not specifically listed under any trait; however the descriptors for Agreeableness fit best (consideration and awareness of others). Unfortunately both communication and influencing others were listed under Extraversion, making it impossible to separate out those two dimensions; therefore no summary values were presented for those dimensions. Refer to Table 3 for the summary statistic data. (For Hypothesis 1c the focus is on MTHM correlations, so the first column, labeled MTHM, is the relevant one.) Using the three benchmark studies, the High grouping of the present study did produce higher MTHM correlations than the benchmark studies found; however the Low grouping demonstrated higher MTHM correlations as well.

In summary, the results of both the relative and absolute analyses failed to support Hypothesis 1c. The dimensions that were rated as consistently providing a

higher opportunity to observe should have yielded higher MTHM correlations compared to those dimensions that were not rated as such. This pattern did not hold for this study.

Hypothesis 1d

Hypothesis 1d stated that if the number of displayed behaviors for a dimension is larger for all relevant exercises, internal convergent validity (using the MTMM framework) will be higher for that dimension, in comparison to those dimensions for which a smaller number of behaviors is consistently displayed and those dimensions for which the number of behaviors displayed is inconsistent across exercises. Therefore, a key aspect of this hypothesis is that one grouping consists of dimensions that display a consistently high number of behaviors across all relevant exercises.

Hypothesis 1d was analyzed by using the average behavior counts to determine the relative levels of opportunity to observe for dimensions. None of the five dimensions displayed a large number of behaviors across all relevant exercises. Oral communication, team building, problem solving, and organizing and planning all were inconsistent across exercises, sometimes yielding a relatively high number of behaviors and sometimes yielding a relatively low number. Influencing others displayed a relatively small number of behaviors for both exercises in which it was observed. Table 7 summarizes these results. Overall, the data did not meet the requirement of consistency stated in the hypothesis, and therefore Hypothesis 1d could not be analyzed further.

Table 7
Average Behavior Counts for Dimensions by Exercise

	Oral Communication	Influencing Others	Team Building	Problem Solving	Organizing & Planning
Resource Allocation	13.92	7.58	7.16	7.50	5.74
In-basket (with interview)	8.15		12.24	13.84	10.31
Policy Analysis	10.31	6.34	7.38	7.81	4.30
Mean behavior count across exercises	10.81	6.96	8.94	9.73	6.81

Note. Influencing others was not assessed for the in-basket exercise, therefore there is no behavior count provided for that combination.

Hypothesis 2: Evaluation Method

Hypothesis 2 stated that the within-dimension evaluation method will lead to more positive construct-related validity results than the within-exercise evaluation method. The present study's assessment center provided a unique opportunity to look at the relative value of the two evaluation methods within the same assessment center.

To test Hypothesis 2 I looked at the construct-related validity evidence produced by two groups of assessors: the primary assessors and the secondary assessors. For these analyses, the primary assessors are those assessors who were specifically assigned to a participant for a certain exercise. They watched the participant, recorded behaviors for the participant, and then made dimension ratings immediately after the completion of the

exercise. The secondary assessors made dimension ratings for the participant only after all exercises were completed, and ratings were made across exercises for each dimension. Thus, the primary assessors represent the within-exercise evaluation method and the secondary assessors represent the within-dimension evaluation method. For each participant there was only one primary assessor rating for each dimension/exercise pairing; however, there were multiple secondary assessors for each. An average across the relevant secondary assessors was used.

Support for Hypothesis 2 would be found if the construct-related validity evidence for the secondary assessors' ratings is more positive than the construct-related validity evidence found for the primary assessors' ratings. The construct-related validity of the two groups was assessed first using Campbell and Fiske's (1959) three conditions for a measure to show construct-related validity: MTHM correlations need to be statistically significant and large enough to suggest convergent validity; MTHM correlations need to be relatively larger than HTHM correlations; and MTHM correlations need to be relatively larger than HTMM correlations. The last two conditions provide evidence of discriminant validity. Campbell and Fiske also asserted that one probably cannot assess absolute validity but should focus on relative validity, specifically stating that validity is demonstrated to the extent that the MTHM correlations are "higher than the average HTHM values" (p. 88).

First consider the evidence for primary assessors (see Appendix E). All but one of the MTHM correlations reached statistical significance ($p < .01$). The correlations

ranged from 0.18 to 0.50 with an overall average MTHM correlation of 0.36. This satisfied the first condition posited by Campbell and Fiske (1959).

The second criterion called for MTHM correlations that are larger than the HTHM correlations. The HTHM correlations for primary assessors ranged from 0.07 to 0.45 and yielded a mean correlation of 0.25. It appeared that the second criterion also had been met in that .36 is larger than .25. However, comparing the two average correlations using the Fisher Z transformation did not yield a statistically significant difference (z = .97, *ns*). Additionally, the lower end of the range of MTHM correlations (.18) fell below the average HTHM correlation (.25), which does not fit with the quoted criterion above.

To further examine the relationship between MTHM and HTHM correlations, and thus provide discriminant validity evidence, the test for differences between dependent correlations (Cohen & Cohen, 1983; Myers & Well, 1991; Steiger, 1980) was utilized. Ninety-six MTHM/HTHM comparisons were made and the pattern of results was studied. (Because of the large number of comparisons a stringent alpha level of .001 was used to offset Type I error.) The overall pattern of results indicated minimal evidence of discriminant validity. Only four of the 96 comparisons yielded statistically significant results in the intended direction (MTHM greater than HTHM). For 12 comparisons the HTHM correlation exceeded the MTMM correlation, but none of these reached statistical significance.

The third criterion offered by Campbell and Fiske called for a comparison of MTHM correlations to HTMM correlations. The average HTMM correlation was .45

with a range of .26 to .60. These values appeared to be higher than the values found for MTHM, thus on the surface this criterion was not met. Further, only one of the 96 MTHM/HTMM comparisons reached statistical significance, and it was in the unintended direction (HTMM greater than MTHM). Overall, the evidence of construct-related validity (as assessed using Campbell and Fiske's criteria) for primary assessors was poor.

The MTMM matrix for secondary assessors was analyzed in the same way. (See Appendix F for MTMM matrix for secondary assessors.) The MTHM correlations were statistically significantly different from zero and were large, ranging from .42 to .73. The average correlation was .57. As with the primary assessors, the first condition presented by Campbell and Fiske was met. For secondary assessors, the HTHM correlations ranged from .16 to .60 and produced an average correlation of .38. The .57 mean MTHM correlation appeared to be higher than the .38 mean HTHM correlation, providing evidence of discriminant validity. This was supported by the Fisher Z transformation and comparison which indicated a statistically significant difference between the two values ($z = 2.03$, $p < .05$). Further, the correlation value that represented the lower range of MTHM correlations was higher than the average HTHM correlation (.42 versus .38), satisfying the quoted Campbell and Fiske criterion.

As with for the primary assessor analysis, the statistical test of the difference between two dependent correlations was used to further examine the relationship between the MTHM and HTHM correlations. The pattern of results yielded positive discriminant validity evidence. Twenty-nine of the 96 comparisons were statistically

significant at the .001 level and all were in the intended direction (MTHM greater than HTHM). Only one comparison was in the opposite direction, but it failed to reach statistical significance (t = -0.22, *ns*).

Table 8
Construct-Related Validity Evidence for Primary versus Secondary Assessors Using Campbell and Fiske's (1959) Criteria

Construct-Related Validity Evidence	Primary Assessors	Secondary Assessors
MTHM correlations		
Mean	.36	.57
Range	.18 - .50	.42 - .73
HTHM correlations		
Mean	.25	.38*
Range	.07 - .45	.16 - .60
HTMM correlations		
Mean	.45	.53
Range	.26 - .60	.31 - .65
Number of MTHM HTHM comparisons that were statistically significant in the *intended* direction	4	29
Number of MTHM HTHM comparisons that were statistically significant in the *unintended* direction	0	0
Number of MTHM HTMM comparisons that were statistically significant in the *intended* direction	0	8
Number of MTHM HTMM comparisons that were statistically significant in the *unintended* direction	1	0

Note. N = 135-138 for Primary Assessors; N = 136-139 for Secondary Assessors.
*Compared to MTHM, $p < .05$.

Looking at the third Campbell and Fiske criterion, the average HTMM correlation for secondary assessors was .53 with a range of .31 to .65. The average MTHM was greater than that for HTMM (.57 versus .53), although the difference was slight and statistically nonsignificant ($z = .54$). Examining the dependent correlation comparison data, eight comparisons were statistically significant at the .001 level and all in the intended direction (MTHM greater than HTMM). Further, only 29 of the 96 comparisons were in the unintended direction (HTMM greater than MTHM), and none of them reached statistical significance ($p > .001$). Overall, the evidence of construct-related validity (as assessed using Campbell and Fiske's criteria) for secondary assessors was positive, especially in comparison to the evidence found for primary assessors.

Table 8 summarizes the construct-related validity evidence for primary and secondary assessors using the criteria set out by Campbell and Fiske (1959). Looking at these data together one can see that the MTHM correlations were higher for secondary assessors versus primary assessors (.57 versus .36). The difference between the MTHM correlations and HTHM correlations also was greater for secondary assessors. Although the average HTHM correlation was smaller for primary assessors (.25 versus .38), it is the relative difference between the MTHM and HTHM correlations that is most important. Finally, the secondary assessor MTMM matrix met the condition of having greater MTHM correlations versus HTMM correlations, whereas the primary assessor matrix did not.

The dependent correlation comparison data further supported this finding. The overall pattern of comparisons for primary assessors showed very few MTHM correlations greater than HTHM correlations and none greater than the HTMM correlations. In fact, the only statistically significant comparison was in the opposite direction, with the HTMM correlation being greater than the MTHM correlation (t = -4.94, $p < .001$). On the other hand, the overall pattern of comparisons for secondary assessors showed a larger number of both HTHM/MTHM and HTMM/MTHM comparisons reaching statistical significance in the intended direction, and no comparison reaching statistical significance in the unintended direction. Therefore, looking at the Campbell and Fiske criteria overall, secondary assessors produced relatively more favorable construct-related validity evidence versus primary assessors, which is consistent with the hypothesis.

As an alternate method of examining discriminant validity, Campbell and Fiske (1959) suggested using a one-tailed sign test (see also McGarty & Smithson, 2005). This binomial sign test involves simply counting the number of HTMM (or HTHM) correlations that are greater than the corresponding MTHM correlation. The null hypothesis is that half of the correlations would be higher just by chance. In other words, discriminant validity would be evidenced by less than half of the HTMM correlations being greater in size than the corresponding MTHM correlation. Note that these HTMM/MTHM comparisons used the same correlation comparisons as were used in the dependent correlations comparison data above; however, the values are used

differently. The HTMM/MTHM comparison results for both primary and secondary assessors are reported in Table 9.

Table 9
Sign Test Results of HTMM versus MTHM Correlations for Primary and Secondary Assessors

	Primary Assessors		
	Resource Allocation/ In-basket Number of HTMM correlations higher than the MTHM correlation (Total possible = 7)	Resource Allocation/ Policy Analysis Number of HTMM correlations higher than the MTHM correlation (Total possible = 8)	In-basket/ Policy Analysis Number of HTMM correlations higher than the MTHM correlation (Total possible = 7)
Oral Communication	5	4	2
Influencing Others		4	
Team Building	7	2	7
Problem Solving	7	8	7
Organizing and Planning	6	8	6
	Secondary Assessors		
Oral Communication	0	0	0
Influencing Others		0	
Team Building	5	1	3
Problem Solving	6	2	1
Organizing and Planning	6	0	5

Note. N = 135-138 for Primary Assessors; N = 136-139 for Secondary Assessors. Influencing others was not assessed for the in-basket exercise, therefore there is no data provided for that combination.

Although neither matrix produced perfect results, the discriminant validity for the secondary assessors did appear to be better. Only two of the 13 comparison groupings for primary assessors had fewer than half of the HTMM correlations smaller than the corresponding MTHM correlation (team building with resource allocation and policy analysis, and oral communication with in-basket and policy analysis). In contrast, nine of the comparison groupings for the secondary assessors met the criterion. Additionally, for six of the 13 comparisons for primary assessors, all of the HTMM correlations were greater than the MTHM correlation, while none of the comparisons for secondary assessors showed all of the HTMM correlations greater than the MTHM correlation.

Aggregating across all comparison groupings, there were 96 HTMM/MTHM comparisons made. For primary assessors, only 23 of the 96 comparisons were in the expected direction, with the normal approximation of the binomial test yielding $z = -5.10$, $p < .01$; meaning that there were significantly fewer comparisons in the wanted direction than expected. Conversely, 67 of the 96 comparisons for secondary assessors were in the expected direction ($z = 3.88$, $p < .01$); indicating that a statistically significant number of comparisons were in the wanted direction.

Table 10 presents the data for the HTHM/MTHM sign test comparisons. Both matrices yielded positive results with no comparison groupings demonstrating more than half of the HTHM correlations being greater than the MTHM. Secondary assessors fared slightly better in that only one comparison grouping had any HTHM correlations greater than the MTMM correlation (problem solving with resource allocation and

in-basket). Contrast that with seven comparison groupings for primary assessors having at least one HTHM greater than the MTMM correlation.

Table 10
Sign Test Results of HTHM versus MTHM Correlations for Primary and Secondary Assessors

	Primary Assessors		
	Resource Allocation/ In-basket	Resource Allocation/ Policy Analysis	In-basket/ Policy Analysis
	Number of HTHM correlations higher than the MTHM correlation (Total possible = 7)	Number of HTHM correlations higher than the MTHM correlation (Total possible = 8)	Number of HTHM correlations higher than the MTHM correlation (Total possible = 7)
Oral Communication	1	0	0
Influencing Others		0	
Team Building	2	0	0
Problem Solving	3	2	2
Organizing and Planning	0	1	1
	Secondary Assessors		
Oral Communication	0	0	0
Influencing Others		0	
Team Building	0	0	0
Problem Solving	1	0	0
Organizing and Planning	0	0	0

Note. N = 135-138 for Primary Assessors; N = 136-139 for Secondary Assessors. Influencing others was not assessed for the in-basket exercise, therefore there is no data provided for that combination.

Overall, the data from both the Campbell and Fiske (1959) criteria and the sign test showed that secondary assessors produced more positive construct-related validity evidence than did primary assessors. Secondary assessors represented the within-dimension evaluation method while primary assessors represented the within-exercise evaluation method. Therefore, Hypothesis 2 was supported in that the within-dimension evaluation method led to more positive construct-related validity results than did the within-exercise evaluation method.

Hypothesis 3: Relationship of Exercises and Dimensions to Performance

Hypothesis 3 stated that the dimension/performance relationships will be stronger than the exercise/performance relationships. In order to test this hypothesis, correlational analyses were performed using the overall effectiveness (OE) measure of performance for the five rating sources: self, subordinate, supervisor, peer, and instructor (Table 11). Dimension and exercise scores were derived by calculating an average across assessors. Both dimensions and exercises yielded the greatest number of statistically significant correlations with self-ratings and peer ratings of overall effectiveness. No dimension or exercise score produced a statistically significant correlation ($p < .05$) with either subordinate ratings or supervisor ratings, and the majority of subordinate correlations were in the negative direction. Only one dimension, oral communication ($r = .28$, $p < .05$), and one exercise, policy analysis ($r = .28$, $p < .05$), resulted in significant correlations with the mean instructor rating; however, the overall effect sizes for the instructor correlations were relatively high (*r*s ranged from .10 to .28).

Table 11
Zero-order Correlations for Dimension and Exercise Scores with Performance

	Performance Ratings – Overall Effectiveness						
	Self-Rating	Mean Subordinate Rating	Mean Supervisor Rating	Mean Peer Rating	Mean Instructor Rating	Mean Scale Score	Standard Deviation
Dimensions							
Oral Communication	.37** (*n* = 110)	-.05 (*n* = 29)	.07 (*n* = 65)	.21* (*n* = 110)	.28* (*n* = 60)	4.42	0.95
Influencing Others	.30** (*n* = 110)	-.08 (*n* = 29)	.09 (*n* = 65)	.16 (*n* = 110)	.10 (*n* = 60)	4.54	1.14
Team Building	.10 (*n* = 110)	.01 (*n* = 29)	.14 (*n* = 65)	.21* (*n* = 110)	.20 (*n* = 60)	4.57	0.97
Problem Solving	.17 (*n* = 110)	-.34 (*n* = 29)	-.05 (*n* = 65)	.19* (*n* = 110)	.24 (*n* = 60)	4.51	0.82
Organizing and Planning	.19* (*n* = 110)	-.11 (*n* = 29)	.11 (*n* = 65)	.21* (*n* = 110)	.18 (*n* = 60)	4.21	0.92
Exercises							
Resource Allocation	.27** (*n* = 111)	-.15 (*n* = 29)	.09 (*n* = 65)	.15 (*n* = 111)	.22 (*n* = 60)	4.36	0.89
In-basket	.25** (*n* = 113)	-.23 (*n* = 29)	.10 (*n* = 65)	.20* (*n* = 113)	.16 (*n* = 61)	4.61	0.85
Policy Analysis	.21* (*n* = 112)	-.01 (*n* = 29)	.06 (*n* = 65)	.26** (*n* = 112)	.28* (*n* = 61)	4.36	0.92
Mean Scale Score	4.17	4.52	4.39	3.95	3.97		
Standard Deviation	0.50	0.38	0.52	0.55	0.53		

Note. OE = overall effectiveness. OE scale ranged from 1 to 5. Dimension and exercise scales ranged from 1 to 7.
* $p < .05$; ** $p < .01$.

Overall, the individual dimension and exercise scores did not demonstrate strong criterion-related validity making it difficult to compare the two groupings of relationships. However, some assessments could be made. Looking at statistical significance, 32% of the dimension/performance relationships and 40% of the exercise/performance relationships reached statistical significance at the .05 level. Using Hemphill's (2003) guidelines for low, medium, and high effect sizes, and considering only those correlations that are in the positive direction, 60% of the dimension/performance relationships were classified as "small," 35% as "medium," and 5% as "large." For the exercise/performance relationships, 42% of the correlations were classified as "small" and 58% as "medium." The strongest relationship was found for a dimension – oral communication with self-ratings of OE (r = .37). In sum, it is unclear from the correlational data alone which facet of the assessment center, exercises or dimensions, has the stronger relationship with overall effectiveness.

Normally, in order to examine the incremental validity of exercises over dimensions a hierarchical regression analysis would be conducted. However, the dimensions were highly intercorrelated, as were the exercises, (see Table 12) making any multiple regression analyses potentially uninterpretable. In an exploratory examination, regressing exercises and dimensions onto supervisor ratings yielded tolerance levels that were unacceptable (less than .000000000001). Therefore, no further regression analyses were conducted or evaluated. In summary, the results indicated little support for Hypothesis 3 as neither dimensions nor exercises exhibited much criterion-related validity with the measures of overall performance in this study.

Table 12
Dimension Intercorrelations

	OC	IO	TB	PS	OP	RA	IB	PA
Oral Communication								
Influencing Others	.703							
Team Building	.643	.573						
Problem Solving	.526	.570	.533					
Organizing & Planning	.561	.569	.604	.600				
Resource Allocation	.807	.840	.714	.660	.705			
In-basket (with interview)	.626	.457	.682	.691	.742	.558		
Policy Analysis	.763	.804	.782	.706	.714	.777	.612	

Note. $N = 136$. OC=oral communication, IO=influencing others, TB=team building, PS=problem solving, OP=organizing and planning, RA=resource allocation, IB=in-basket (with interview), PA=policy analysis. All correlations statistically significant at $p < .01$.

Stability of Dimensions across Exercises: Exploratory

This exploratory expectation stated that the more stable dimensions (oral communication and problem solving), those that are more similar to traits, will have higher MTHM correlations than the more situationally specific dimensions (team building, and organizing and planning) which are expected to demonstrate lower MTHM correlations, and thus, lower internal construct-related validity. To test this I calculated the MTHM correlations for the two separate groups. As shown in Table 13, the more

stable group demonstrated higher MTHM correlations than the less stable group for two of the three exercise pairings (resource allocation with in-basket and in-basket with policy analysis), as well as the average across exercise pairings. This provided initial support for the exploratory investigation. However, none of the pairings resulted in statistically significant differences (Steiger, 1980); and using Fisher Z transformations and confidence intervals (Myers & Well, 1991) did not yield a statistically significant difference for the overall average ($z = 0.38$, *ns*; lower bound of confidence interval around oral communication and problem solving average correlation = .43 and upper bound = .57). These findings did not demonstrate strong support for the exploratory analyses - there were no statistically significant differences between the dimensions that were perceived as more stable and those that were more situationally specific.

Table 13
Monotrait-Heteromethod Correlations for Groupings Based on Stability of Dimensions across Exercises

Exercise Pairings	**More Stable Dimensions** Oral Communication and Problem Solving	**Less Stable Dimensions** Team Building, and Organizing & Planning
Resource Allocation – In-basket	0.43	0.41
Resource Allocation – Policy Analysis	0.58	0.61
In-basket – Policy Analysis	0.50	0.44
Average[a]	0.50	0.48

Note. $k = 424$ to 429 for both groups, where k is the number of data points.
[a] Indicates average correlation across all exercise pairings.

Although the analysis was stated in relative terms, the data also were analyzed using the absolute benchmarks from Hypothesis 1 - Hemphill's (2003) guidelines for effect sizes, and the three assessment center construct-related validity articles that provided summary statistics for MTMM matrices (Bowler & Woehr, 2006; Jones, 1992; Lievens et al., 2006). Using Hemphill's benchmark, both groupings (more and less stable) demonstrated "high" correlations. This did not fit with the expected result that the stable dimensions would yield high MTHM correlations and the less stable dimensions would yield lower MTHM correlations. Table 3 provided the summary statistics for the three assessment center MTMM benchmark studies; the MTHM column was the column of interest for these analyses. All correlations for this exploratory analysis were higher than all relevant correlations in the benchmark studies, again indicating a lack of support for the hypothesis.

Taken together, the relative and absolute analyses produced results that generally did not support the exploratory investigation. The dimensions that were more stable (oral communication and problem solving) did not produce larger MTHM correlations than the dimensions that were less stable (team building, and organizing and planning).

CONCLUSIONS AND DISCUSSION

The objective of the present study was to provide more evidence with respect to the assessment center construct-related validity debate. Specifically four pieces of the puzzle were examined by taking a closer look at opportunity to behave and observe, evaluation method, the relationship to criterion-related validity, and the stability of dimensions. It was posited that the so-called exercise effect found in many assessment center construct-related validity studies is artifactual and not real.

There was some support for the hypotheses overall. Specifically, Hypothesis 1, involving the opportunity to behave within exercises, was supported when the rank ordering was determined by actual behavior counts as opposed to assessor ratings. When operationalized this way, discriminant validity was higher for those exercises that provided a greater opportunity to behave. Therefore, as some researchers have suggested (e.g., Brannick et al., 1989; Harris, Becker, & Smith, 1993; Highhouse & Harris, 1993; Joyce et al., 1994; Kleinmann & Koller, 1997; Reilly et al., 1990) the low construct-related validity demonstrated in other assessment centers may be due to a small number of behaviors being displayed by participants, and therefore a small number of behaviors on which to base ratings.

Evaluation method, addressed with Hypothesis 2, also provided some positive construct-related validity evidence. Secondary assessors, who represented the within-dimension method, demonstrated better overall construct-related validity than did primary assessors, who represented the within-exercise method. Although lab studies have reached similar conclusions (Robie et al., 2000; Woehr & Arthur, 2003), applied

studies have produced mixed results. The present study adds a new contribution in that both evaluation methods, within-exercise and within-dimension, were examined within the same assessment center with the same participants and assessors. Given this more direct comparison, it is notable that the results are favorable to the within-dimension method.

Limited support was found for the exploratory hypothesis that examined the effect of the stability of dimensions on the construct-related validity of assessment centers. The convergent validity of the more stable dimensions (oral communication and problem solving) was greater than the convergent validity found for the less stable dimensions (team building, and organizing and planning), but the differences between the two groupings were not large or statistically significant.

Some interesting results were found when I examined the criterion-related validity for exercises and dimensions separately (Hypothesis 3). There was low criterion-related validity overall for both sets of predictors; correlations ranged from .01 to .37 (absolute value) and there was no consistent pattern to the strength or statistical significance of the correlations for exercises versus dimensions. Due to high predictor intercorrelations, regression analyses were not able to be properly conducted. The overall lack of positive results may lead back to the idea that we need to move away from OARs and focus on exercise- and dimension-level data within the assessment center (Arthur et al., 2003); perhaps the same can be said for performance ratings. The criterion used in this study was an overall effectiveness rating. So although the predictor data were examined at the exercise and dimension level, it was not possible to relate

those directly to both exercise- and dimension-level performance criteria. This could account for the low levels of criterion-related validity found.

Another interesting finding was the difference in rank ordering that the two methods of measuring opportunity to behave yielded (Hypotheses 1a and 1b). The assessor ratings and the behavior counts yielded opposite rank ordering – the two LGDs were classified as providing higher opportunity to behave by the assessors but behavior counts indicated the opposite, while the in-basket exercise was rated as providing lower opportunity to behave by the assessors but yielded high average behavior counts. This discrepancy may be explained by the objectivity of the behavior counts versus the assessor ratings. Further, assessors may have felt that the two LGDs provided more opportunities because of the verbal nature of the exercises, as opposed to the in-basket which had a large written component.

The analyses for the opportunity to observe for dimensions (Hypotheses 1c and 1d) did not yield any noteworthy results. Using assessor ratings as the ranking variable and then assessing the convergent validity of the AC did not produce the expected result. Further, the behavior counts did not supply the data necessary to create the high and low groupings – there were no consistently high behavior counts for any one dimension – therefore, H1d was not able to be assessed at all.

Taken together, the results of this study suggest that there *are* some areas of design and implementation that can affect the construct-related validity of assessment centers. The convergent and divergent evidence found for the within-dimension evaluation method and the discriminant validity found for those exercises that presented

enough behaviors for assessors to make appropriate ratings are strong enough to encourage researchers to continue to investigate ways to improve assessment center validity. This is in contrast to those who have called for a stop to this search in favor of looking at assessment centers as work samples only (e.g., Lance et al., 2004; Lance, 2008b). Although much of the past research has found a stronger exercise effect versus dimension effect, there are still a number of factors that have not been examined thoroughly and to a resounding conclusion. The addition of this study should remind researchers that the debate about the construct-related validity of assessment centers is still relevant today.

Limitations and Suggestions for Future Research

Clearly, no single study can resolve the issue of construct-related validity within assessment centers. However, the present study is: a new empirical examination of two ideas presented under the umbrella of design and implementation flaws being the reason for the discrepancies in construct-related validity (opportunity to behave and observe, and evaluation method); an investigation of the relationship of criterion-related validity within the context of construct-related validity for the same assessment center; and a beginning step into the domain of dimensions as personality variables. Although many of the results were positive, there are some issues that should be noted.

The behavior counts used in Hypotheses 1b and 1d provided a unique challenge in that they are, obviously, not a precise measure of opportunity to observe and behave. One specific limitation is that there may have been more behaviors listed for the in-basket exercise simply because there was more time for assessors to transcribe

behaviors. For the two LGDs, assessors made most of their written observations during the session, while usually focused on two participants at a time. For the in-basket, assessors went through each participant's written notations individually; and although there were time constraints on the review process, it was not done in real-time. Given these potential limitations, the present study does still suggest that when participants provide enough dimension-related behaviors on which assessors can base their ratings, the discriminant validity of the assessment center can be relatively high. Thus, opportunity to behave within exercises does appear to affect the construct-related validity of this assessment center.

Another potential limitation of the present study is that the primary and secondary assessors' ratings are necessarily confounded (Hypothesis 2). Although the secondary assessors are present during the LGDs, their ratings are based mainly on the behavioral information reported by the primary assessor. However, despite the fact that secondary assessors base their ratings to a large extent on what behaviors the primary assessors provide them, secondary assessors still demonstrated more positive construct-related validity evidence than the primary assessors. Thus, the within-dimension effect was strong enough to overcome the restriction of receiving information from an assessor who was rating using the within-exercise evaluation method.

Lastly, the changing nature of the assessment center may be a limitation. Throughout the duration of the study, the assessment center had changes made to it because of various factors such as time constraints, changes in what the client was

looking to assess, and discovering aspects of past assessment centers that did not work as well as intended. For example, dimensions were added and then subtracted (e.g., innovation); an exercise was added to specifically assess negotiation; and the exercise content changed. This adapting process is not new to those who have worked with applied assessment centers – they seem to always be a work in progress, evolving to fit the demands of the client, the job, the situation. However, to the extent that assessment centers in applied settings can be the same, these administrations are.

Many of the posited reasons for why assessment centers have demonstrated low construct-related validity have assumed that the lack of convergent and discriminant validity is an error of some sort; be it an error in the way the assessment center was designed and implemented, an error in the statistical procedures used to analyze the evidence, or an error in the construct specified. They all have assumed that the desired, indeed the true, outcome is consistency in ratings within dimension across exercises (e.g., Kolk et al., 2001). But perhaps cross-situational consistency is not the true picture of what is occurring in an assessment center; maybe instead there is actual variation in performance across exercises. It seems plausible that a person's performance can vary from situation to situation; for example, that one can have good oral communication skills one-on-one, but poor oral communication skills with a group. It seems plausible that a person may understand how to build-up team members in a group setting, but that team building is manifested differently when one has to do so in a memo. The idea that we would expect behavior to be consistent across exercises, much less ratings of said behavior, ignores most personality and performance appraisal literature.

The fundamental question of whether assessment centers are measuring the dimensions they are supposed to be measuring does not necessarily need to be answered by looking at cross-situational consistency. In other words, do the dimension ratings have to be consistent across exercises in order for an assessment center to have construct-related validity? It seems not. If the dimension ratings were completely consistent across exercises then we would not need more than one exercise to measure the dimension. It could be measured just as well with only one exercise. It seems that one objective of having different exercises is to get at different aspects of the same dimension - tap into different parts. Therefore, should researchers even be looking for cross-situational consistency within assessment centers? With assessment centers, designers and assessors are looking at the whole picture - what happens across exercises. (This is reflected in the ratings – dimension ratings and OARs are used.) A dimension can be measured accurately across exercises and not necessarily be consistent across exercises (Lievens, 2001b; Lievens, 2002). This possibly makes dimensions less like traits than some researchers may have been assuming.

Perhaps that is the problem with using MTMM matrices to assess the construct-related validity of assessment centers – the method leads one to believe that we are looking at traits that are supposed to be consistent across situations. However, we can be measuring a dimension accurately and still not obtain cross-situational consistency because part of the definition of the dimension is that it may be manifested differently in different situations. Assessment centers often demand assessors to look at the dimension across various situations in order to find the "typical" performance of the

person - not the "average." This is a similar issue that occurs in performance ratings (e.g., Woehr & Miller, 1997). Do we want the person who is consistently average or the person who is great on some things and horrible on others so that they average out to be "average?" It is a dilemma. One that leads me to believe that MTMM is not the best way to examine the construct-related validity of assessment centers (see also the "Why assessment centers do not work the way they are supposed to" focal article and commentaries in *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 2008, 1(1); Lance, Woehr, & Meade, 2007). Other methods of assessing internal construct-related validity should be sought after and considered.

Along similar lines, attention should be paid to how dimension definitions are written - performance based (good to have always; more is better) or trait-based (a certain amount is good, but too much or too little may not be). Scale anchors and interpretation of the anchors also should be considered. There could be confusion over absolute versus appropriateness. For example, some assessors may view the highest rating on a scale as being representative of a person possessing or exhibiting the most of the dimension possible (absolute). Other assessors may view the highest end of the scale as representing a person who exhibited the appropriate amount of a dimension, which may or may not be a "large" amount of the dimension (appropriateness). Taking these two things into consideration may help researchers decide whether MTMM is a viable and appropriate method for assessing the internal construct-related validity of assessment centers. They also may shed some light on why we are getting differences in the evidence we are finding.

Finally, to say that we have exhausted the search for characteristics that affect the construct-related validity of assessment centers after only two decades of examination seems a bit premature. As the introduction points out, there are numerous attributes that have yet to be studied thoroughly and yet to be resolved. Future research should continue to examine the effects of various design and implementation factors that affect assessment center construct-related validity, but should also look for ways other than MTMM to assess said validity.

REFERENCES

Adler, S. (1987). Toward the more efficient use of assessment center technology in personnel selection. *Journal of Business and Psychology*, *2*, 74-93.

Ahmed, Y., Payne, T., & Whiddett, S. (1997). A process for assessment exercise design: A model of best practice. *International Journal of Selection and Assessment, 5*, 62-68.

Arthur, W., Jr. (1997). *Development and implementation of the Texas State Auditor's Office management development center*. The Texas State Auditor's Office contract. Bryan, TX: Winfred Arthur, Jr. Consulting.

Arthur, W., Jr. (2001). *Development and implementation of the Bush School leadership skills assessment and development program*. The Bush School of Government and Public Service contract. Bryan, TX: Winfred Arthur, Jr. Consulting.

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*, 125-154.

Arthur, W., Jr., Day, E. A., & Woehr, D. J. (2008). Mend it, don't end it: An alternate view of assessment center construct-related validity evidence. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 105-111.

Arthur, W., Jr., & Tubre, T. (2002). *The assessment center construct-related validity paradox: A case of construct misspecificaton?* Unpublished paper, Texas A&M University, College Station, TX.

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

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.

Baron, H., & Janman, K. (1996). Fairness in the assessment center. In C. L. Cooper and I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (Vol. 11, pp. 61-114). Chichester, England: John Wiley & Sons Ltd.

Binning, J. F., Adorno, A. J., & LeBreton, J. M. (1999). *"Sociotechnical" moderators of assessment center criterion-related validity.* Paper presented at the Fourteenth Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478-494.

Borman, W. C. (1982). Validity of behavioral assessment for predicting military recruiter performance. *Journal of Applied Psychology, 67*, 3-9.

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.

Brannick, M. T., Michaels, C. E., & Baker, D. P. (1989). Construct validity of in-basket scores. *Journal of Applied Psychology, 74,* 957-963.

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.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Chan, D. (1996). Criterion and construct validation of an assessment centre. *Journal of Occupational and Organizational Psychology, 69,* 167-181.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences (*2nd edition*)*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Connelly, B. S., Ones, D. S., Ramesh, A. & Goff, M. (2008). A pragmatic view of assessment center exercises and dimensions. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 121-124.

Crawley, B., Pinder, R., & Herriot, P. (1990). Assessment centre dimensions, personality and aptitudes. *Journal of Occupational Psychology, 63,* 211-216.

Dean, M.A., Roth, P. L., & Bobko, P. (2008). Ethnic and gender subgroup differences in assessment center ratings: A meta-analysis. *Journal of Applied Psychology, 93,* 685-691.

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

Dreher, G. F., & Sackett, P. R. (1981). Some problems with applying content validity evidence to assessment center procedures. *Academy of Management Review, 6*, 551-560.

Dugan, B. (1988). Effects of assessor training on information use. *Journal of Applied Psychology, 73*, 743-748.

Dulewicz, V. (1991). Improving assessment centres. *Personnel Management, 23*, 50-55.

Fitzgerald, L. F., & Quaintance, M. K. (1982). Survey of assessment center use in state and local government. *Journal of Assessment Center Technology, 5*, 9-21.

Fleenor, J. W. (1996). Constructs and developmental assessment centers: Further troubling empirical findings. *Journal of Business and Psychology, 10*, 319-335.

Fletcher, C. A., & Dulewicz, V. (1984). An empirical study of a U.K.-based assessment centre. *Journal of Management Studies, 21*, 83-97.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493-511.

Gaugler, B. B., & Thornton, G. C. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74*, 611-618.

Haaland, S., & Christiansen, N. D. (2002). Implications of trait-activation theory for evaluating the construct validity of assessment center ratings. *Personnel Psychology, 55*, 137-163.

Harris, M. M., Becker, A. S., & Smith, D. E. (1993). Does the assessment center scoring method affect the cross-situational consistency of ratings? *Journal of Applied Psychology, 78*, 675-678.

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*, 78-80.

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*, 140-155.

Hoeft, S., & Schuler, H. (2001). The conceptual basis of assessment centre ratings. *International Journal of Selection and Assessment, 9*, 114-123.

Howard, A. (1997). A reassessment of assessment centers: Challenges for the 21st century. *Journal of Social Behavior and Personality, 12,* 13-52.

Howard, A. (2008). Making assessment centers work the way they are supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 98-104.

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.

Joiner, D. A. (2000). Guidelines and ethical considerations for assessment center operations: International Task Force on Assessment Center Guidelines. *Public Personnel Management, 29,* 315-331.

Jones, R. G. (1992). Construct validation of assessment center final dimension ratings: Definition and measurement issues. *Human Resource Management Review, 2*, 195-220.

Jones, R. G. (1997). A person perception explanation for validation evidence from assessment centers. *Journal of Social Behavior and Personality, 12*, 169-178.

Jones, R.G., & Klimoski, R. J. (2008). Narrow standards for efficacy and the research playground: Why either-or conclusions do not help. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 137-139.

Joyce, L. W., Thayer, P. W., & Pond, S. B. (1994). Managerial functions: An alternative to traditional assessment center dimensions? *Personnel Psychology, 47*, 109-121.

Kauffman, J. R., Jex, S. M., Love, K. G., & Libkuman, T. M. (1993). The construct validity of assessment centre performance dimensions. *International Journal of Selection and Assessment, 4*, 213-223.

Kleinmann, M. (1993). Are rating dimensions in assessment centers transparent for participants? Consequences for criterion and construct validity. *Journal of Applied Psychology, 78*, 988-993.

Kleinmann, M., & Koller, O. (1997). Construct validity of assessment centers: Appropriate use of confirmatory factor analysis and suitable construction principles. *Journal of Social Behavior and Personality, 12*, 65-84.

Kleinmann, M., Kuptsch, C., & Koller, O. (1996). Transparency: A necessary requirement for the construct validity of assessment centres. *Applied Psychology: An International Review, 45*, 67-84.

Klimoski, R., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology, 40*, 243-260.

Kolk, N. J., Born, M. P., & van der Flier, H. (2001). *Common rater variance as explanation for the lack of construct validity of assessment center dimensions*. Paper presented at the Sixteenth Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Krause, D. E., Kersting, M., Heggestad, E. D., & Thornton, G. C., III. (2006). Incremental validity of assessment center ratings over cognitive ability tests: A study at the executive management level. *International Journal of Selection and Assessment, 14*, 360-371.

Kudisch, J. D., Ladd, R. T., & Dobbins, G. H. (1997). New evidence on the construct validity of diagnostic assessment centers: The findings may not be so troubling after all. *Journal of Social Behavior and Personality, 12*, 129-144.

Kuptsch, C., Kleinmann, M., & Koller, O. (1998). The chameleon effect in assessment centers: The influence of cross-situational behavioral consistency on the convergent validity of assessment centers. *Journal of Social Behavior and Personality, 13*, 102-116.

Lance, C. E. (2008a). Where have we been, how did we get there, and where shall we go? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 140-146.

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

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.

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.

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.

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.

Landy, F. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*, 1183-1192.

Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection and Assessment, 6*, 141-152.

Lievens, F. (2001a). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86*, 255-264.

Lievens, F. (2001b). Assessors and use of assessment centre dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior, 22*, 203-221.

Lievens, F. (2002). Trying to understand the different pieces of the construct validity puzzle of assessment centers: An examination of assessor and assessee effects. *Journal of Applied Psychology, 87*, 675-686.

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.

Lievens, F., & Conway, J. M. (2000). *Analysis of multitrait-multimethod data in assessment centers: Methodological and substantive issues.* Paper presented at the Fifteenth Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

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.

Lievens, F., & Keer, E. V. (1999). *Modeling method effects in assessment centers: An application of the correlated uniqueness approach*. Paper presented at the

Fourteenth Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Lievens, F., & Keer, E. V. (2001). The construct validity of a Belgian assessment centre: A comparison of different models. *Journal of Occupational and Organizational Psychology, 74*, 373-378.

Macan, T. H., Avedon, M. J., Paese, M., & Smith, D. E. (1994). The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology, 47*, 715-738.

McEvoy, G. M., & Beatty, R. W. (1989). Assessment centers and subordinate appraisals of managers: A seven-year examination of predictive validity. *Personnel Psychology, 42*, 37-52.

McEvoy, G. M., Beatty, R. W., & Bernardin, H. J. (1987). Unanswered questions in assessment center research. *Journal of Business and Psychology, 2*, 97-111.

McGarty, C., & Smithson, M. (2005). Independence and nonindependence: A simple method for comparing groups using multiple measures and the binomial test. *European Journal of Social Psychology, 35*, 171-180.

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 dimension ratings. *Journal of Applied Psychology, 93,* 1042-1052.

Myers, J. L., & Well, A. D. (1991). *Research design and statistical analysis.* New York, NY: HarperCollins Publishers Inc.

Neidig, R. D., Martin, J. C., & Yates, R. E. (1979). The contribution of exercise skill ratings to final assessment center evaluations. *Journal of Assessment Center Technology, 2*, 21-33.

Neidig, R. D., & Neidig, P. J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology, 69*, 182-186.

Norton, S. D. (1977). The empirical and content validity of assessment centers vs. traditional methods for predicting managerial success. *Academy of Management Review, 2*, 442-453.

Norton, S. D. (1981). The assessment center process and content validity: A reply to Sackett and Dreher. *Academy of Management Review, 6*, 561-566.

Pynes, J., & Bernardin, H. J. (1992). Mechanical vs consensus-derived assessment center ratings: A comparison of job performance validities. *Public Personnel Management, 21*, 17-28.

Reilly, R. R., Henry, S., & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology, 43*, 71-84.

Ritchie, R. J., & Moses, J. L. (1983). Assessment center correlates of women's advancement into middle management: A 7-year longitudinal analysis. *Journal of Applied Psychology, 68*, 227-231.

Robertson, I., Gratton, L., & Sharpley, D. (1987). The psychometric properties and design of managerial assessment centres: Dimensions into exercises won't go. *Journal of Occupational Psychology, 60*, 187-195.

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.

Russell, C. J. (1985). Individual decision processes in an assessment center. *Journal of Applied Psychology, 70*, 737-746.

Russell, C. J. (1987). Person characteristic versus role congruency explanations for assessment center ratings. *Academy of Management Journal, 30*, 817-826.

Russell, C. J., & Domm, D. R. (1995). Two field tests of an explanation of assessment centre validity. *Journal of Occupational and Organizational Psychology, 68*, 25-47.

Sackett, P. R. (1982). A critical look at some common beliefs about assessment centers. *Public Personnel Management, 11*, 140-147.

Sackett, P. R. (1987). Assessment centers and content validity: Some neglected issues. *Personnel Psychology, 40*, 13-25.

Sackett, P. R., & Dreher, G. F. (1981). Some misconceptions about content-oriented validation: A rejoinder to Norton. *Academy of Management Review, 6*, 567-568.

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67*, 401-410.

Sackett, P. R., & Dreher, G. F. (1984). Situation specificity of behavior and assessment center validation strategies: A rejoinder to Neidig and Neidig. *Journal of Applied Psychology, 69*, 187-190.

Sackett, P. R., & Harris, M. M. (1988). A further examination of the constructs underlying assessment center ratings. *Journal of Business and Psychology, 3*, 214-229.

Sagie, A., & Magnezy, R. (1997). Assessor type, number of distinguishable dimension categories, and assessment centre construct validity. *Journal of Occupational and Organizational Psychology, 70*, 103-108.

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (1999). *A new frame for frame of reference training: Enhancing the construct validity of assessment centers*. Paper presented at the Fourteenth Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407-422.

Schmitt, H., & Noe, R. A. (1983). Demonstration of content validity: Assessment center example. *Journal of Assessment Center Technology, 6*, 5-11.

Schmitt, N., & Ostroff, C. (1986). Operationalizing the "behavioral consistency" approach: Selection test development based on a content-oriented strategy. *Personnel Psychology, 39*, 91-108.

Schmitt, N., Schneider, J. R., & Cohen, S. A. (1990). Factors affecting validity of a regionally administered assessment center. *Personnel Psychology, 43*, 1-12.

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.

Shore, T. H., Shore, L. M., & Thornton, G. C. (1992). Construct validity of self- and peer evaluations of performance dimensions in an assessment center. *Journal of Applied Psychology, 77*, 42-54.

Shore, T. H., Thornton, G. C., & Shore, L. M. (1990). Construct validity of two categories of assessment center dimension ratings. *Personnel Psychology, 43*, 101-116.

Silverman, W. H., Dalessio, A., Woods, S. B., & Johnson, R. L. (1986). Influence of assessment center methods on assessors' ratings. *Personnel Psychology, 39*, 565-578.

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). College Park, MD: Author.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251.

Tenopyr, M. L. (1977). Content-construct confusion. *Personnel Psychology, 30*, 47-54.

Thornton, G. C., III, Tziner, A., Dahan, M., Clevenger, J. P., & Meir, E. (1997). Construct validity of assessment center judgments: Analyses of the behavioral reporting method. *Journal of Social Behavior and Personality, 12*, 109-128.

Turnage, J. J., & Muchinsky, P. M. (1982). Transsituational variability in human performance within assessment centers. *Organizational Behavior and Human Decision Processes, 30,* 174-200.

Turnage, J. J., & Muchinsky, P. M. (1984). A comparison of the predictive validity of assessment center evaluations versus traditional measures in forecasting supervisory job performance: Interpretive implications of criterion distortion for the assessment paradigm. *Journal of Applied Psychology, 69,* 595-602.

Tziner, A., & Dolan, S. (1982). Validity of an assessment center for identifying future female officers in the military. *Journal of Applied Psychology, 67,* 728-736.

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 Management, 29,* 231-258.

Woehr, D. J., & Miller, M. J. (1997). Distributional ratings of performance: More evidence for a new rating format. *Journal of Management, 23,* 705-720.

APPENDIX A

DIMENSION DEFINITIONS

Oral Communication: The extent to which an individual effectively conveys oral information and responds to questions and challenges.

Influencing Others: The extent to which an individual is effective in persuading others to do something or adopt a point of view in order to produce desired results without creating hostility.

Team Building: The extent to which an individual successfully engages and works in collaboration with members of a group such that others are involved in, and contribute to, the process and outcome.

Problem Solving: The extent to which an individual gathers data; effectively analyzes and uses data and information; generates viable options, ideas, and solutions; selects supportable courses of action for problems and situations; generates new or creative ideas and solutions; and uses available resources in new and more effective ways.

Organizing and Planning: The extent to which an individual effectively and systematically arranges his/her own work and resources as well as that of others for efficient task accomplishment; and the extent to which the individual anticipates and prepares for the future.

APPENDIX B

BEHAVIORAL ANCHORS FOR DIMENSIONS

ORAL COMMUNICATION: The extent to which an individual effectively conveys oral information and responds to questions and challenges.

Clearly below expected: 1

The participant was unable to communicate ideas effectively. Presentations were delivered in a manner that was not logically thought out and organized. Eye contact was not made with group members and the pitch and delivery rate of the messages were inappropriate. This participant failed to listen and respond appropriately to questions.

Expected: 4

The participant offered well formed and logical presentations, spoke in a clear and concise manner, and made eye contact with those to whom speaking. The participant paid attention to others, listened to others, and facilitated the integration of new information into ideas and suggestions from others.

Clearly above expected: 7

The participant's ideas were presented in an easy-to-follow format that was well formed and logically arranged. Ideas were clear and the participant held listener's interest through the effective use of voice, gestures, and other visual aides. The participant used eye contact and active listening. This participant properly targeted messages to his/her audience and was able to evoke strong feelings in listeners.

INFLUENCING OTHERS: The extent to which an individual is effective in persuading others to do something or adopt a point of view in order to produce desired results without creating hostility.

Clearly below expected: 1

The participant had no impact on the outcome or direction the group took. The participant did not generate support for his/her own or others' positions. The participant failed to generate discussion, bring out others' ideas, or counter others' proposals, and may have had a strong negative effect on group progress. The participant did not motivate others to alter behaviors or ideas in any way.

Expected: 4

The participant generated some support and discussion for a position and attempted to direct the group's movement toward accepting a position. The participant also provided some information and assistance in an attempt to alter the ideas and behaviors of others. The participant elicited ideas from others in an attempt to influence the interaction.

Clearly above expected: 7

The participant successfully gained support for either their, or another participant's position, and greatly affected the direction the group took. The participant had a great deal of influence on the adoption of a position that was acceptable to most without creating a hostile environment. The participant motivated others to alter behavior and ideas, and brought others into discussions in order to move the group forward. The participant unified the group through effecting compromise.

TEAM BUILDING: The extent to which an individual successfully engages and works in collaboration with members of a group such that others are involved in, and contribute to the process and outcome.

Clearly below expected: 1

The participant worked independently of the other group members and argued for their own position excluding others' input. The participant failed to recognize the validity of the views and opinions of others and failed to reinforce and reward subordinates for their efforts.

Expected: 4

The participant helped clarify group goals, worked to support other's views, and solicited input and feedback from subordinates and peers. The participant reinforced and rewarded the suggestions and efforts of subordinates and peers. The participant identified and took advantage of opportunities to delegate activities to subordinates and peers.

Clearly above expected: 7

The participant worked with other group members to clarify group goals, solicited input and feedback from subordinates and peers, acknowledged and defended alternative viewpoints, and engaged in supportive interaction with other members to resolve conflict and arrive at consensus. The participant sought opportunities to appropriately delegate activities to and reinforce and reward subordinates and peers.

PROBLEM SOLVING: The extent to which an individual gathers data; effectively analyzes and uses data and information; generates viable options, ideas, and solutions; selects supportable courses of action for problems and situations; generates new or creative ideas and solutions; and uses available resources in new and more effective ways.

Clearly below expected: 1
The participant made little use of available information, failed to attend to details, and did not gather input from other sources to analyze information and develop solutions. The participant failed to demonstrate logic behind assumptions, and did not consider alternative solutions or the effects of constraints. The participant neglected to explore the source, role, and causes of the problem in developing alternatives. The participant failed to choose a clear course of action. * The participant was constrained by boundaries in generating solutions to the problems that were presented. The argumentation and approaches used by the participant were not original. Real life examples were not supplied by the participant. The participant did not recognize or support innovative thinking in others.

Expected: 4
The participant used most of the available information and attended to the most salient details to identify surface problems. Some information was gathered from other sources and some tradeoffs were identified. The effect of constraints was considered, and the short–term impact of solutions was weighed. The participant evaluated alternative solutions and suggested a clear course of action. * The participant used some creative argumentation to support proposals. By reordering a problem or going outside imposed boundaries, the participant generated some new ideas or solutions to problems. The participant recognized and supported innovative thinking in others. The participant proposed some original solutions or new approaches.

Clearly above expected: 7
The participant effectively used available information and gathered new and relevant data to explore the underlying issues related to the problem. A number of feasible alternatives were developed and corresponding tradeoffs were identified. The constraints were adequately considered, and both the long– and short–term effects of possible solutions were evaluated. There was a clear, logical relationship between the analysis that was made of the problem and the decisions that the participant selected. A clear course of action was chosen. * The participant redefined the problem using creative argumentation, real life examples, or ideas from others so that new options could be considered. Frequently, the participant was the first to suggest new ideas to the group to help move past imposed constraints. The participant used creative parallels to put familiar things together in an unusual but effective way. By seeing and supporting or stimulating innovative thinking in others, the participant contributed to the generation of breakthrough solutions.

ORGANIZING AND PLANNING: The extent to which an individual effectively and systematically arranges his/her own work and resources as well as that of others for efficient task accomplishment; and the extent to which the individual anticipates and prepares for the future.

Clearly below expected: 1

The participant did not prioritize or categorize activities. Their level of pacing resulted in failure to accomplish important tasks. The participant failed to use outlines, charts, or lists and focused only on short-term goals. The participant handled each item separately and did not recognize relationships between tasks or potential problems. The participant failed to delegate or coordinate diverse actions, people, or events. The participant also failed to provide instructions, deadlines, or follow-up plans.

Expected: 4

The participant prioritized or outlined some of the task completion process and considered some long-range goals. The participant paced himself/herself to accomplish some important tasks. The participant grouped tasks and scheduled activities, but not in great detail. The participant was able, to some extent, to coordinate individuals or events and recognize scheduling conflicts. The participant frequently provided clear instructions and used appropriate resources when delegating tasks.

Clearly above expected: 7

The participant arranged their and others' resources and work in order to accomplish most tasks efficiently. The participant grouped tasks, prepared agendas, recognized high and low priority tasks, and kept a schedule of activities with deadlines. The participant recognized relationships between tasks, potential problems, and necessary resources before taking action. The participant organized the work of others by delegating tasks to the appropriate people, providing clear instructions, deadlines, and evaluating results. The participant effectively integrated long- and short-term goals.

APPENDIX C

ASSESSOR MEASURE OF OPPORTUNITY TO OBSERVE

Dear Assessor:

Thank you for taking the time to complete this measure. The results of this measure will be used in aggregate form for my dissertation analyses. Overall, I am examining the construct-related validity of assessment centers. However, the specific objective of this measure is to obtain information on the varying levels of opportunity to observe for each dimension/exercise combination.

Once you have read the directions and reviewed the exercises and dimension definitions, **the measure should take only about 5 minutes to complete**. If you have any questions about the measure or how the results will be used, please feel free to contact me.

Again, thank you so much for your help!

Measure of Opportunity to Observe within an Assessment Center

Directions:
For each exercise/dimension combination, please rate the level of opportunity to observe on a scale of 1 to 7 (scale defined below). Opportunity to observe is defined here as the extent to which the dimension behaviors can be displayed within the exercise. For example, think about the dimension of organizing and planning within the in-basket exercise. If you feel that participants have a great deal of opportunity to display organizing and planning behaviors within the in-basket exercise, then you would rate this combination a 7. If, on the other hand, you think there is no opportunity to display pertinent behaviors, then you would rate this combination a 1. (Note that I do not expect "high" ratings simply by virtue of the fact that I am asking you to provide ratings.)

After the five dimension ratings for each exercise, there is a Summary Ratings page asking you to rate the overall level of opportunity to observe for each exercise. For this rating please consider the exercise across all five dimensions: how well does the exercise elicit behaviors for all dimensions?

Scale:
1 = **No Opportunity to Observe**; there were no behaviors within this dimension that this exercise could elicit
2
3
4 = **Moderate Opportunity to Observe**; there were a moderate number of behaviors within this dimension that this exercise could elicit
5
6
7 = **Great Deal of Opportunity to Observe**; there were a large number of behaviors within this dimension that this exercise could elicit

The following two pages provide summaries of the three exercises and definitions of the five dimensions. Please review this information before continuing on to complete the measure.

EXERCISE SUMMARIES:

Resource Allocation Exercise (competitive leaderless group discussion)

This exercise involved the participants role playing executive directors of government bureaus of a hypothetical country, Simlandia. The participants were to reach a group consensus on how to best allocate an unexpected surplus of money. The goal of each participant was to obtain as much money as possible for his/her own bureau, as well as aid the group in making the best overall decision about the allocation. Participants had 30 minutes to review materials and come up with proposals. Each participant made a 5 minute presentation and then the group had 50 minutes to reach an agreement.

In-Basket Exercise

This exercise involved the participant assuming the duties of the general manager for the Bradford Consolidated Fund. The participant had three hours to go through items in the general manager's in-basket and take any action deemed necessary. Assessors then reviewed the materials and actions taken by the participant, and conducted an in-basket interview to clarify any discrepancies and obtain information on the participant's rationale for his/her responses to items.

Policy Analysis Exercise (non-competitive leaderless group discussion)

For this exercise participants acted as members of a team of consultants asked to give recommendations to a client concerning a management problem. The "problem" for most assessment centers was safety rule compliance and involved office workers in a factory not wearing their hard hats when they walked through the production area. [The "problem" for the Bush School 1997 assessment center involved deciding whether or not to attract a professional sports franchise to the city.] The team was to discuss the problem and come to an agreement on the most appropriate solution. Participants had 15 minutes to study the problem and come up with their own recommendations. Then the group had 50 minutes to discuss the ideas and reach an agreement.

DIMENSION DEFINITIONS:

Oral Communication: The extent to which an individual effectively conveys oral information and responds to questions and challenges.

Influencing Others: The extent to which an individual is effective in persuading others to do something or adopt a point of view in order to produce desired results without creating hostility.

Organizing and Planning: The extent to which an individual effectively and systematically arranges his/her own work and resources as well as that of others for efficient task accomplishment; and the extent to which the individual anticipates and prepares for the future.

Problem Solving: The extent to which an individual gathers data; effectively analyzes and uses data and information; generates viable options, ideas, and solutions; selects supportable courses of action for problems and situations; generates new or creative ideas and solutions; and uses available resources in new and more effective ways.

Team Building: The extent to which an individual successfully engages and works in collaboration with members of a group such that others are involved in, and contribute to, the process and outcome.

Review

If you wish, you may review the information on the Exercise Summaries or the Dimension Definitions by clicking on the "back" button on your browser to go back to the appropriate page.

When you feel like you have refamiliarized yourself with the exercises and dimensions, click on the "Next" button below to continue on and complete the measure.

NOTE: Unfortunately, you will not be able to change your responses once you submit your answers for each page. Please consider your answers carefully before clicking the "Next" button on each of the following pages. (This is a software limitation.)

Resource Allocation Exercise (competitive leaderless group discussion)

This exercise involved the participants role playing executive directors of government bureaus of a hypothetical country, Simlandia. The participants were to reach a group consensus on how to best allocate an unexpected surplus of money. The goal of each participant was to obtain as much money as possible for his/her own bureau, as well as aid the group in making the best overall decision about the allocation. Participants had 30 minutes to review materials and come up with proposals. Each participant made a 5 minute presentation and then the group had 50 minutes to reach an agreement.

	1 No Opportunity to Observe	2	3	4 Moderate Opportunity to Observe	5	6	7 Great Deal of Opportunity to Observe
Oral Communication: The extent to which an individual effectively conveys oral information and responds to questions and challenges.	○	○	○	○	○	○	○
Influencing Others: The extent to which an individual is effective in persuading others to do something or adopt a point of view in order to produce desired results without creating hostility.	○	○	○	○	○	○	○
Organizing and Planning: The extent to which an individual effectively and systematically arranges his/her own work and resources as well as that of others for efficient task accomplishment; and the extent to which the individual anticipates and prepares for the future.	○	○	○	○	○	○	○
Problem Solving: The extent to which an individual gathers data; effectively analyzes and uses data and information; generates viable options, ideas, and solutions; selects supportable courses of action for problems and situations; generates new or creative ideas and solutions; and uses available resources in new and more effective ways.	○	○	○	○	○	○	○
Team Building: The extent to which an individual successfully engages and works in collaboration with members of a group such that others are involved in, and contribute to, the process and outcome.	○	○	○	○	○	○	○

In-Basket Exercise

This exercise involved the participant assuming the duties of the general manager for the Bradford Consolidated Fund. The participant had three hours to go through items in the general manager's in-basket and take any action deemed necessary. Assessors then reviewed the materials and actions taken by the participant, and conducted an in-basket interview to clarify any discrepancies and obtain information on the participant's rationale for his/her responses to items.

	1 No Opportunity to Observe	2	3	4 Moderate Opportunity to Observe	5	6	7 Great Deal of Opportunity to Observe
Oral Communication: The extent to which an individual effectively conveys oral information and responds to questions and challenges.	○	○	○	○	○	○	○
Influencing Others: The extent to which an individual is effective in persuading others to do something or adopt a point of view in order to produce desired results without creating hostility.	○	○	○	○	○	○	○
Organizing and Planning: The extent to which an individual effectively and systematically arranges his/her own work and resources as well as that of others for efficient task accomplishment; and the extent to which the individual anticipates and prepares for the future.	○	○	○	○	○	○	○
Problem Solving: The extent to which an individual gathers data; effectively analyzes and uses data and information; generates viable options, ideas, and solutions; selects supportable courses of action for problems and situations; generates new or creative ideas and solutions; and uses available resources in new and more effective ways.	○	○	○	○	○	○	○
Team Building: The extent to which an individual successfully engages and works in collaboration with members of a group such that others are involved in, and contribute to, the process and outcome.	○	○	○	○	○	○	○

Policy Analysis Exercise (non-competitive leaderless group discussion)

For this exercise participants acted as members of a team of consultants asked to give recommendations to a client concerning a management problem. The "problem" for most assessment centers was safety rule compliance and involved office workers in a factory not wearing their hard hats when they walked through the production area. (The "problem" for the Bush School 1997 assessment center involved deciding whether or not to attract a professional sports franchise to the city.) The team was to discuss the problem and come to an agreement on the most appropriate solution. Participants had 15 minutes to study the problem and come up with their own recommendations. Then the group had 50 minutes to discuss the ideas and reach an agreement.

	1 No Opportunity to Observe	2	3	4 Moderate Opportunity to Observe	5	6	7 Great Deal of Opportunity to Observe
Oral Communication: The extent to which an individual effectively conveys oral information and responds to questions and challenges.	○	○	○	○	○	○	○
Influencing Others: The extent to which an individual is effective in persuading others to do something or adopt a point of view in order to produce desired results without creating hostility.	○	○	○	○	○	○	○
Organizing and Planning: The extent to which an individual effectively and systematically arranges his/her own work and resources as well as that of others for efficient task accomplishment; and the extent to which the individual anticipates and prepares for the future.	○	○	○	○	○	○	○
Problem Solving: The extent to which an individual gathers data; effectively analyzes and uses data and information; generates viable options, ideas, and solutions; selects supportable courses of action for problems and situations; generates new or creative ideas and solutions; and uses available resources in new and more effective ways.	○	○	○	○	○	○	○
Team Building: The extent to which an individual successfully engages and works in collaboration with members of a group such that others are involved in, and contribute to, the process and outcome.	○	○	○	○	○	○	○

Summary Ratings

Please rate each exercise on its overall ability to elicit dimension-related behaviors. In other words, overall, how well does the exercise elicit behaviors for all dimensions in general?

	1 Provides No Opportunity to Observe	2	3	4 Provides Moderate Opportunity to Observe	5	6	7 Provides a Great Deal of Opportunity to Observe
Resource Allocation Exercise: Competitive leaderless group discussion.							
In-Basket Exercise							
Policy Analysis Exercise: Non-Competitive leaderless group discussion							

APPENDIX D

LINKAGES BETWEEN DIMENSIONS AND EXERCISES

Assessment Center Dimensions	EXERCISES		
	Resource Allocation (Competitive Leaderless Group Discussion)	In-Basket	Policy Analysis (Non-Competitive Leaderless Group Discussion)
Oral Communication	X	X	X
Influencing Others	X		X
Team Building	X	X	X
Problem Solving	X	X	X
Organizing and Planning	X	X	X

Note: Shaded areas represent dimensions that are unobservable in the specified exercise.

APPENDIX E

MTMM MATRIX FOR PRIMARY ASSESSORS

	OC-RA	OC-IB	OC-PA	IO-RA	IO-PA	TB-RA	TB-IB	TB-PA	PS-RA	PS-IB	PS-PA	OP-RA	OP-IB	OP-PA
OC-RA	--													
OC-IB	.38**	--												
OC-PA	.47**	.48**	--											
IO-RA	.56**	.42**	.45**	--										
IO-PA	.36**	.37**	.52**	.50**	--									
TB-RA	.44**	.27**	.41**	.43**	.26**	--								
TB-IB	.17	.41**	.33**	.28**	.18*	.24**	--							
TB-PA	.22*	.31**	.50**	.32**	.46**	.50**	.38**	--						
PS-RA	.42**	.36**	.33**	.61**	.36**	.42**	.20*	.24**	--					
PS-IB	.14	.29**	.25**	.28**	.18*	.09	.38**	.14	.18*	--				
PS-PA	.12	.21*	.33**	.23**	.44**	.29**	.29**	.43**	.33**	.23**	--			
OP-RA	.48**	.30**	.44**	.50**	.31**	.51**	.23**	.33**	.49**	.18*	.28**	--		
OP-IB	.18*	.26**	.27**	.12	.16	.12	.47**	.19*	.07	.43**	.22**	.39**	--	
OP-PA	.32**	.29**	.40**	.31**	.54**	.28**	.18*	.46**	.25**	.17*	.46**	.35**	.28**	--

Note. N = 135-138. OC = oral communication; IO = influencing others; TB = team building; PS = problem solving; OP = organizing & planning; RA = resource allocation exercise; IB = in-basket exercise; PA = policy analysis exercise
* $p < .05$; ** $p < .01$.

APPENDIX F

MTMM MATRIX FOR SECONDARY ASSESSORS

	OC-RA	OC-IB	OC-PA	IO-RA	IO-PA	TB-RA	TB-IB	TB-PA	PS-RA	PS-IB	PS-PA	OP-RA	OP-IB	OP-PA
OC-RA	--													
OC-IB	.61**	--												
OC-PA	.73**	.68**	--											
IO-RA	.60**	.45**	.60**	--										
IO-PA	.60**	.46**	.63**	.67**	--									
TB-RA	.56**	.42**	.58**	.61**	.48**	--								
TB-IB	.26**	.41**	.46**	.32**	.30**	.45**	--							
TB-PA	.46**	.49**	.65**	.48**	.55**	.64**	.54**	--						
PS-RA	.53**	.44**	.47**	.59**	.52**	.49**	.22*	.42**	--					
PS-IB	.26**	.37**	.35**	.30**	.32**	.24**	.45**	.28**	.42**	--				
PS-PA	.31**	.31**	.41**	.33**	.49**	.29**	.35**	.51**	.57**	.52**	--			
OP-RA	.57**	.43**	.48**	.58**	.53**	.59**	.33**	.45**	.60**	.40**	.38**	--		
OP-IB	.26**	.31**	.30**	.29**	.26**	.26**	.52**	.32**	.16	.58**	.32**	.45**	--	
OP-PA	.52**	.36**	.50**	.48**	.59**	.50**	.30**	.57**	.47**	.36**	.45**	.70**	.47**	--

Note. N = 135-138. OC = oral communication; IO = influencing others; TB = team building; PS = problem solving; OP = organizing & planning; RA = resource allocation exercise; IB = in-basket exercise; PA = policy analysis exercise
* $p < .05$; ** $p < .01$.

VITA

NAME: **Kathryn Diane Archuleta**

ADDRESS: Texas A&M University – Department of Psychology
College Station, TX 77843-4235

EDUCATION

M.S. Texas A&M University, Industrial/Organizational Psychology, 1998.

B.A. Rice University, Psychology and Managerial Studies, 1995.

PROFESSIONAL EMPLOYMENT/EXPERIENCE

July 2001 to November 2004 — **Jeanneret & Associates, Inc.**, Houston, TX. INTERN. Responsibilities include performing job analyses, developing and validating selection systems, conducting statistical analyses, writing reports, maintaining databases, serving as interviewer and assessor for client, coaching client personnel, and scoring selection and promotion tests.

September 1999 to May 2001 — **Department of Management**, Texas A&M University. LECTURER. Responsibilities included teaching upper-level undergraduate Human Resource Management class, preparing and presenting multi-media lectures, and maintaining class website.

PUBLICATIONS/PRESENTATIONS

Arthur, W., Jr., Tubre, T. C., Day, E., Sheehan, M. K., Sanchez- Ku, M. L., Paul, D. S., Paulus, L. E., & Archuleta, K. D. (2001). Motor vehicle crash involvement and moving violations: Convergence of self-report and archival data. *Human Factors*, 43, 1-11.