ASSESSMENT CENTER CONSTRUCT VALIDITY: METHODOLOGICAL AND DESIGN MODIFICATIONS

by

ALLISON BARI SIMINOVSKY

(Under the Direction of Charles E. Lance)

ABSTRACT

Assessment centers (ACs) have remained a popular selection and development tool for years. Although ACs often demonstrate criterion-related validity, researchers have struggled with finding content validity for AC dimensions. The problem exists for two reasons: nonconvergence and inadmissibility issues when using the MTMM framework, and dominance of dimension variance by exercise variance. The current study reexamined previously reported AC matrices, reclassified using three different schemes of broad dimension factors in order to increase each model's indicator-to-factor ratio to promote convergence and admissibility. Additionally, a number of design modifications were examined as to their influence on increasing dimension variance. Results show a significant increase in convergence and admissibility rates as indicator-to-factor ratio increases. However, the remaining analyses reached inconclusive and nonsignificant results. Possible explanations for these results are discussed, as well as ramifications for the findings.

INDEX WORDS: Assessment Center, MTMM, Construct Validity, Indicator-to-Factor Ratio, Dimension Factors

ASSESSMENT CENTER CONSTRUCT VALIDITY: METHODOLOGICAL AND DESIGN MODIFICATIONS

by

ALLISON BARI SIMINOVSKY

B.B.A., CUNY Bernard M. Baruch College, 2010

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2013

ASSESSMENT CENTER CONSTRUCT VALIDITY: METHODOLOGICAL AND DESIGN MODIFICATIONS

by

ALLISON BARI SIMINOVSKY

Major Professor: Charles E. Lance

Committee: Nathan T. Carter
Robert P. Mahan

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2013

DEDICATION

For J.Z.H.

ACKNOWLEDGEMENTS

I wish to proffer countless millennia of gratitude and admiration to Chuck Lance, for his seemingly infinite patience and constant support and assistance on this project. I also express sincere gratitude to the other members of my thesis committee, Nathan Carter and Rob Mahan, for their thoughtful and constructive feedback and enthusiasm. I give the utmost appreciation and respect to Brian Hoffman for his vision and support. Finally, I offer my thanks to Julia Sauer, Stefanie Beck, and Nadia van Hauwaert for their assistance in the translation of several articles used in this study.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

CHAPTER 1

INTRODUCTION

Assessment centers (ACs) have remained a popular selection and training tool for decades as they provide candidates with the opportunity to demonstrate job-related behaviors in a setting similar to that of the actual workplace (Gaugler, Rosenthal, Thornton, & Bentson, 1987; Thornton & Byham, 1982). ACs were conceptualized to examine an individual's standing on a number of constructs measured via a series of exercises simulating on-the-job tasks (Thornton & Byham, 1982). For example, a participant could be rated on dimensions such as consideration for others, verbal communication, and persuasiveness in a leaderless group discussion exercise. In a typical AC, raters score participants on several dimensions at the end of each exercise (post exercise dimension ratings, or PEDRs; Lance, 2008). Practitioners and researchers alike value this tool for its realistic simulations of job tasks (Klimoski & Brickner, 1987) and its ability to predict job performance and other work-related outcomes, such as promotability (Chan, 1996) and potential for future performance (Thornton & Gibbons, 2009). ACs routinely demonstrate criterion-related validity for the prediction of performance (Arthur, Day, & McNelly, 2003; Gaugler et al., 1987; Meriac, Hoffman, Woehr, & Fleisher, 2008). However, the construct validity of ACs remains an unresolved issue, as rated dimensions across exercises very often fail to show convergent validity (Cahoon, Bowler, & Bowler, 2012; Lance, Lambert, Gewin, Lievens, & Conway, 2004). Instead, different dimensions within individual exercises are highly correlated, calling into question the utility of AC dimension measurement (e.g. Klimoski & Brickner, 1987; Lance, 2008).

Researchers often examine AC ratings using the popular multitrait-multimethod design (MTMM; Campbell & Fiske, 1959) analyzed with confirmatory factor analysis (CFA). This framework assumes that each exercise represents a method and each dimension represents a trait. CFA functions to examine the similarity of the fit of the data to that of the theoretical model, and in this case helps to determine whether methods and traits have the same relationship in reality as was theorized (Kline, 2011). ACs were conceptualized to demonstrate correlated ratings for like dimensions across job-related exercises and minimally correlated ratings for unique dimensions within exercises (Sackett & Dreher, 1982). It is typically found in the resulting correlation matrices from CFA, however, that unique dimensions within exercises are correlated and like dimensions across exercises are minimally related, which fails to support the construct validity of ACs (Bycio, Alvares, & Hahn, 1987; Lance, Lambert et al., 2004; Neidig & Neidig, 1984). As a result, AC researchers have placed a heavy focus on methods to fix the construct validity problem.

As ACs are quite complex in design, prior studies have focused on the effects of specific AC characteristics on demonstrating dimension construct validity (Lievens, 1998; Woehr & Arthur, 2003). Some characteristics for which they found significant effects include the number of dimensions rated, the length of rater training, and the occupation of AC raters. It has been speculated that these characteristics play a role in construct validity, likely due to their differential functioning on rater cognitive strain, which is to say that if the characteristics are less prone to cause cognitive strain (i.e. fewer rated dimensions), ratings will be more accurate and researchers will have less difficult in finding construct validity. Unfortunately, results have been inconclusive regarding AC design characteristics, causing many to consider methodological modifications as the potential remedy to the construct validity problem.

Researchers have also focused on various methodological alterations to fix the construct validity problem. Whereas different MTMM structures have been suggested (Hoffman, Melchers, Blair, Kleinmann, & Ladd, 2011; Lance, Lambert et al., 2004; Lievens & Conway, 2001), the correlated dimension-correlated exercise structural model (CDCE) has routinely been found to be the most conceptually appropriate in the study of ACs (Lance, Lambert et al., 2004; Monahan, 2011). Despite its theoretical appropriateness, the CDCE model as applied to ACs often poses problems when analyzed using CFA, most frequently those of model nonconvergence and inadmissibility (Lance, Woehr, & Meade, 2007). As a result, some have called for the elimination of dimensions from AC studies altogether and have instead suggested a focus on exercises only (Jackson, Barney, Stillman, & Kirkley, 2007; Lance, 2008; Neidig & Neidig, 1984). This viewpoint is not popular, likely due to its antithetical stance toward the original purpose of ACs: to measure a participant's standing on various dimensions needed for performance on the job (Bowler & Woehr, 2006; Bowler & Woehr, 2009; Howard, 2008). Despite issues with establishing construct validity, the use of multiple dimensions in ACs has occasionally been supported (Guenole, Chernyshenko, Stark, Cockerill, & Drasgow, 2013; Hoffman et al., 2011). Therefore, it remains necessary to develop a means of demonstrating construct validity in ACs in order to recommend their continued use.

It has recently been proposed that problems with construct validity might relate specifically to an artifact of measurement (Hoffman et al., 2011). Researchers have found that by increasing the indicator-to-factor ratio of an MTMM-like matrix in CFA, one is more likely to obtain a convergent and admissible solution (Tomás, Hontangas, & Oliver, 2000). This could allow for an interpretation of construct validity that supports the continued use of dimensions in ACs (Hoffman et al., 2011; Monahan, 2011). One means of increasing the indicator-to-factor

ratio is the use of broad dimension factors, or catch-all categories that narrow down the list of measurable dimensions into several more concrete categories (Woehr & Arthur, 2003). For example, the unique dimensions of verbal communication, written communication, and responsiveness could all fall into the category of "communication", and therefore serve as indicators of this broader dimension category. By increasing the number of indicators in proportion to the number of factors through the use of broad dimensions, the assessment of construct validity becomes a more straightforward process that eliminates some of the issues that typically accompany the CDCE model in AC studies.

The present study attempts to build upon the results of Hoffman et al. (2011) and Monahan (2011) through the reanalysis of previously reported AC CDCE models. It is expected that when the originally reported dimensions are reclassified into broad dimension factors, this increase in indicator-to-factor ratio will lead to more frequent convergent and admissible solutions in CFA, demonstrating construct validity and subsequently supporting the continued measurement of dimensions in ACs. As this method will alter AC variance components, I will also reexamine the relationship between construct validity and AC design characteristics (Lievens, 1998; Woehr & Arthur, 2003) as to their potential contributions to dimension construct variance with respect to broad dimension factors. If successful in its goals, this study can help to put to rest the debate of construct validity in dimensions vs. exercises in ACs. In order to examine potential solutions to the AC construct validity problem, it is first important to thoroughly examine the nature of the problem and proposed solutions that have failed to work in the past.

CHAPTER 2

LITERATURE REVIEW AND HYPOTHESES

The Construct Validity Paradox

The examination of construct validity has arguably been one of the most popular and controversial courses of study in AC literature, with recent years being no exception to this trend (Arthur, Day, & Woehr, 2008; Bowler & Woehr, 2009; Cahoon et al., 2012). However, researchers have had difficulty in demonstrating that the PEDRs of like constructs have high correlations across exercises (Bycio et al., 1987; Lance, Newbolt, Gatewood, Foster, French, & Smith, 2000). Distinct constructs within exercises, however, are often highly intercorrelated, supporting the content validity of AC exercises above that of constructs. This lack of discriminant validity within exercises calls into question the very nature and purpose of ACs (Lance, 2008).

There have been numerous suggestions as to what would remedy the construct validity problem, including universal clarification of dimensions and exercises (Howard, 2008), use of different structural models (Hoffman et al., 2011; Lievens & Conway, 2001), and overall measurement and design characteristics of ACs (Arthur, Woehr, & Maldegen, 2000; Woehr & Arthur, 2003). However, these efforts have failed to fully remedy the construct validity problem. While there have been calls for the redesign of ACs to measure exercises instead of constructs in order to take advantage of robust exercise effects often present in ACs (Jackson et al., 2007; Lance et al., 2000; Neidig & Neidig, 1984), this idea has not been popular, as most researchers are hesitant to back away from what ACs were designed to measure: dimensions (Lance, 2008).

Numerous suggestions have been offered to improve AC design in order to see dimension effects, such as enhancing assessor training (Lievens, 2001; Schleicher, Day, Mayes, & Riggio, 2002), using fewer rated dimensions (Gaugler & Thornton, 1989; Schneider & Schmitt, 1992), and employing "expert" raters (Lievens, 2001; Sagie & Magnezy, 1997). Although these interventions have resulted in slight improvements in construct validity for dimensions, exercise variance still dominates dimension variance (Bowler & Woehr, 2006; Lance, Lambert et al., 2004). Therefore, academicians have been in an ongoing struggle to explain the challenges behind establishing construct validity for ACs. It is important to consider the unique measurement properties of ACs in order to gain a clear picture of the construct validity problem.

Measurement. Assessment centers are used to examine the pattern of assessees' ratings on dimensions across exercises. This structure is compatible with the idea of a multitrait-multimethod matrix (Campbell & Fiske, 1959; Sackett & Dreher, 1982). Using the Campbell and Fiske (1959) methodology, the matrix is examined to determine *convergent* and *discriminant* validity evidence (Bagozzi & Yi, 1991). In order to establish evidence of convergent validity, the matrix must show significant monotrait-heteromethod correlations, demonstrating that measured traits account for covariance between the corresponding trait-method units. Discriminant validity can be demonstrated in part via relatively low heterotrait-monomethod correlations. Despite the usefulness of this method, it has several problems that are not conducive to its practical use. These problems include the assumption of equal reliability for all measures and the assumption that methods are uncorrelated (Bagozzi & Yi, 1991). Additionally, this method forces the researcher to interpret variance in a subjective manner, meaning interpretations of variance rely on the opinions of the individual conducting the analysis. Overall, the Campbell and Fiske methodology is not suitable for AC analysis.

The CFA method is often used to analyze MTMM matrices and is more appropriate for AC research (Bagozzi & Yi, 1991). AC researchers have often elected to view constructs as traits and exercises as methods, and the correlations in the matrix are those generated from PEDRs (Sackett & Dreher, 1982). A primary benefit of the CFA approach is that it allows one to partition total variance into that attributed by method (i.e. exercise), trait, and error components (Bagozzi & Yi, 1991). This facilitates the examination of construct validity in ACs. Additionally, CFA is a theoretically flexible approach, meaning it is possible for the researcher to examine a variety of theoretical factor structures by changing a model's factor pattern matrix and factor covariance structure (Kleinmann & Köller, 1997).

The CDCE structural model is predominant in AC analysis (Anderson, Lievens, Van Dam, & Born, 2006; Lance, Lambert, et al., 2004; Monahan, 2011). In particular, three reviews have assessed the best-fitting structure for AC analysis. Lievens and Conway (2001) used CFA to examine three models: dimension-only, with correlated dimensions; exercise-only, with correlated exercises; and a combination model which represented exercises as correlated uniquenesses (CU). The CU model was found to have the best fit for the most sets of data. Additionally, it was found that dimension and exercise variance contributions were equal (.34), disconfirming the argument that exercises contribute more to variance than do dimensions.

In another study, Lance, Lambert et al. (2004) drew attention to problems with the CU model, namely that this structure results in inflation of dimension effects, which is misleading to interpretation. Furthermore, Lievens and Conway (2001) included inadmissible solutions in their evaluation of variance. When reanalyzing the same data sets, Lance, Lambert et al. (2004) found that a single dimension, correlated exercise model provided the best overall fit, followed by a

correlated exercise-only model. They then examined the admissible solutions and determined that exercise variance contributions were substantially higher than those of dimensions.

In Bowler and Woehr's (2006) meta-analysis, they recoded relevant studies using Arthur et al.'s (2003) six-dimension taxonomy and Spychalski, Quiñones, Gaugler, and Pohley's (1997) six-exercise taxonomy. They examined six different structural models and found the CDCE model to be most appropriate. Although they found exercises to contribute more variance in indicators than dimensions, the magnitude of the range of variance was less than that found in Lance, Lambert et al. (2004). It is crucial to note that Bowler and Woehr (2006) fixed several exercise intercorrelations to zero in order to circumvent nonconvergence and inadmissibility issues. Such modifications to models should only be made on the basis of theoretical rationale, not statistical convenience (Kline, 2011). Therefore, these results should be interpreted with caution.

Overall, these reviews have inconsistent results with regard to the most appropriate model structure for ACs. There are problems with the CDCE model in achieving model convergence and proper solutions (Dumenci & Yates, 2012; Lance et al., 2007), even though this model is most theoretically compatible with AC structure. As there has not yet been a methodological solution to the AC construct validity problem, researchers have looked elsewhere; in addition to methodological considerations, various AC design characteristics have also been analyzed as to their role in construct validity.

Design characteristics. Various AC characteristics have been assessed in the past in order to determine their role in the contribution of dimension variance (Arthur et al., 2000; Lievens, 1998; Woehr & Arthur, 2003). In Lievens' (1998) review, experimental and pseudo-experimental studies were sorted into five categories based on their source of experimental

manipulation: dimensions, situational exercises, assessor characteristics, systematic observation and evaluation procedures, and integration of results. Differences in dimension, exercise, and assessor characteristics were found to impact construct validity (Lievens, 1998). Woehr & Arthur (2003) examined the history of a wide number of AC design and measurement characteristics in their extensive review and meta-analysis of these factors and their relationship to construct validity with varied results. The various characteristics studied were related to both AC design, as well as administration.

Number of dimensions. It has been repeatedly found that dimension variance is more prevalent when raters are asked to assess fewer dimensions, as well as those dimensions that are squarely conceptually distinct from one another (Lievens, 1998; Russell, 1985; Woehr & Arthur, 2003). Specifically, the aforementioned reviews support using a lower number of dimensions and those dimensions which are conceptually distinct from one another as a means to increase dimension convergent validity (Lievens, 1998; Woehr & Arthur, 2003). These results are in line with the theory that the use of broad dimension factors should increase dimension variance (Hoffman et al., 2011). The use of fewer and more clearly operationalized dimensions is thought to both decrease cognitive load on raters (Gaugler & Thornton, 1989) and to impact convergence and admissibility rates (Hoffman et al., 2011).

Participant-to-assessor ratio. On a similar note, Woehr & Arthur (2003) hypothesized that the participant-to-assessor ratio could impact construct validity in that when asked to rate too many participants, raters might face too heavy of a cognitive burden. Although this hypothesis was not supported, the review was conducted a number of years ago and therefore did not include more recent papers. It may be beneficial to reexamine the effect of participant-to-assessor ratio when taking into account more recent findings.

Rating approach. There are two primary rating approaches most often used in the assignment of PEDRs (Sackett & Dreher, 1982). The within-exercise approach has raters assess participants on dimensions at the end of each individual exercise. The across-exercise approach has raters assign PEDRs at the end of all of the AC exercises. The across-exercise approach has been found to result in higher dimension construct validity, as it allows raters to see a fuller spectrum of participant behavior across any given dimension (Robie, Osburn, Morris, Etchegaray, & Adams, 2000; Woehr & Arthur, 2003).

Rater type. AC administrators frequently make use of two main types of raters: managers and industrial-organizational psychologists (Woehr & Arthur, 2003). Both of these roles require understanding of job behaviors and the necessary traits for good performers. However, it has been found that there is higher dimension construct validity with psychologist raters instead of managers, likely due to the experience and knowledge required by the former group (Sagie & Magnezy, 1987; Woehr & Arthur, 2003).

Training. It is common for AC raters, particularly managers, to undergo training prior to AC administration, and this training has been found to impact construct validity, presumably because those raters who receive training have a clearer and more systematic approach to judging and rating participant behavior than those raters who do not receive training (Woehr & Arthur, 2003). Furthermore those raters who receive training for a longer period of time have increased opportunity to comprehend and practice the skills necessary for rating, further impacting dimension construct validity (Lievens, 1998; Woehr & Arthur, 2003).

Purpose. ACs generally function for the purposes of either selection and promotion (administrative purposes) or training and development. Woehr and Arthur (2003) found that ACs that focused on training and development had higher construct validity because raters are more

cognitively focused on differentiating behaviors when examining participants in the context of development.

***Use of behavioral checklists*.** Behavioral checklists differ from traditional AC rating approaches in that instead of being rated directly on dimensions, applicants are rated in terms of the presence of a number of behaviors (Jackson et al., 2007). Researchers can increase the number of indicators through the use of behavioral checklists, which should theoretically help to find construct validity (Monahan, 2011). In his review, Lievens (1998) found mixed results for the hypothesis that those ACs using behavioral checklists for ratings would demonstrate higher dimension construct validity than those using other rating methods, such as graphical rating scales. However, the increase in indicator-to-factor ratio has been subsequently found improve convergence and admissibility rates (Monahan, 2011).

The aforementioned design characteristics, despite their potential role in resolving the construct validity problem, have not been examined since 2003 on a large scale. As a result, there have been many studies that have gone unexamined as to the impact of AC design characteristics on construct validity. Accordingly, it is time that a new review take place. Furthermore, Monahan (2011) demonstrated that the problems with construct validity appear to be reflections of measurement artifact. In particular, AC convergence problems appear to be at least somewhat concurrent with the use of small indicator-to-factor ratios in models (Tomás et al., 2000). If a solution for measurement issues can be found, perhaps we can achieve more conclusive results with respect to design characteristics.

Indicator-to-factor ratio. It is generally considered a "rule-of-thumb" when conducting a CFA to use as many indicators as possible for each factor, with a practical minimum being three indicators per factor (Kline, 2011). On a practical level, this should ensure adequate

coverage of the domain of the factor in question, assuming that the indicators used have strong psychometric properties (Bandalos & Boehm, 2009). It has also been found that when using multiple indicators per factor, the CDCE model is better fitting and more theoretically appropriate than the CTCU model in that it achieves more stable estimates, and this finding is even more robust when more indicators are used (Tomás et al., 2000). A high indicator-to-factor ratio can also contribute to reaching a model solution. Using a series of Monte Carlo simulations, Marsh, Hau, Balla, and Grayson (1998) demonstrated that higher indicator-to-factor ratios increase one's chances of achieving model convergence and admissibility when using CFA. In addition, it has been found that when the indicator-to-factor ratio is high, there is less need for a large sample size (Marsh et al., 1998; Velicer & Fava, 1998).

Furthermore, they also found that higher indicator-to-factor ratios tended to have more accurate and stable (i.e. lower standard errors) parameter estimates than those models with lower ratios. Bandalos and Boehm (2009) described this phenomenon:

> Just as we need adequate samples of people to approximate population quantities related to characteristics of such people, we need adequate samples of variables to approximate the population quantities related to the variables. Sampling too few variables can result in the same types of instability in estimating variable-related properties as can sampling too few people when estimating population parameters (p. 79).

Hoffman et al. (2011) found increased levels of convergence and admissibility using a CDCE + g hierarchical factor structure, with g representing a general performance factor. This model makes use of broad dimension factors, which are general thematic categories encompassing multiple AC dimensions that can increase the indicator-to-factor ratio when used in lieu of unique micro-dimensions. For example, aggressiveness, need for advancement, and

persistence are examples of specific dimensions that comprise the broad construct "drive" in Arthur et al.'s seven-dimension taxonomy (2003). This approach treats ratings as individual indicators, partially eliminating overlap between micro-level dimensions and allowing some dimensions to correlate more strongly than others (Hoffman et al., 2011). Constructs can be more easily distinguished from one another with the use of broad dimension factors, as very similar constructs instead serve as indicators of the same broad dimension factor. This helps to demonstrate construct variance and provide support for the inclusion of dimensions in ACs. Furthermore, there has been criticism about the lack of rigor in the psychometric analysis of dimensions in AC research (Howard, 2008) which could potentially be resolved with the employment of validated broad dimension factors. It has been suggested that the hybrid CDCE + g model may not be accounting for improved solution rates through its structure, but rather through its increase in indicator-to-factor ratio. Monahan (2011) demonstrated that the CDCE model had the best fit, even over the hybrid model, when higher indicator-to-factor ratios were used. Therefore, it is worthwhile to consider broad dimension factors further.

Broad dimension factors. The broad dimension categories of Arthur et al. (2003) resulted from a meta-analysis of AC dimension criterion-related validity, during which 168 unique dimensions were consolidated into seven broader representative categories. The unique dimensions were first sorted into Thornton and Byham's (1982) 33 commonly used dimensions, and were then further collapsed into a new list of seven overarching factors. Six of these dimension categories demonstrated criterion-related validity[1]. The seven dimension factors are: communication, consideration/awareness of others, drive, influencing others, organizing and planning, problem solving, and tolerance for stress/uncertainty (Arthur et al., 2003).

[1] One of the seven categories, tolerance for stress/uncertainty, was not examined in Arthur et al.'s (2003) study due to a dearth of data points available in the existing literature.

The Borman & Brush (1993) taxonomy arose from the need for clear and complete set of dimensions for the evaluation of managerial performance. In order for the resulting dimensions to be applicable, an effort was made to choose factors that were behaviorally relevant, stemming from critical incidents, instead of those theorized by researchers. This resulted in 246 empirical dimensions. These dimensions were evaluated and sorted, eventually resulting in 4 "mega-dimensions" which were found to hold up with previous taxonomies of managerial performance. These dimensions are: technical activities/mechanics of management, leadership and supervision, communication and interpersonal facilitation, and other useful behavior.

Prior to the aforementioned classification schemes, Shore et al. (1990) devised a system of two broad dimension factors: interpersonal-style and performance-style categories. They used eleven unique dimensions that fit into these categories in their examination of an AC. Such classification schemes stem from the need to identify observable behaviors that encompass the two main tasks of leaders: initiating structure and consideration (House, Filley, & Kerr, 1971). Initiating structure refers to the task-oriented aspects of leadership, such as organization and problem solving; consideration consists of expressive and emotional aspects of leadership (House et al., 1971). The influence of these two categories is clearly visible in the various schemes of broad AC dimension factors (see Table 2.1).The results indicated that cognitive ability test scores were significantly correlated with performance dimensions, but not interpersonal dimensions, showing convergent and discriminant validity for the broad dimension factors. Overall, the use of broad dimension factors was found to be beneficial to the demonstration of construct validity in this study.

The three aforementioned classification schemes, despite their differing numbers of factors, have been found to encompass the same general pattern of AC dimensions (leadership

behaviors, task-oriented behaviors and communication-oriented behaviors) and are viewed as nested within one another, as demonstrated in Table 2.1 (Hoffman et al., 2011). When using broad dimension factors, one can eliminate overlap between similar dimensions: similar dimensions are interpreted as indicators of the same broad dimension factor, which increases the indicator-to-factor ratio (Hoffman et al., 2011). Thus, it is expected that the broader the taxonomy used, the larger the increase in indicator-to-factor ratio will be.

The Present Study

In order to build upon the findings of Hoffman et al. (2011) and Monahan (2011), my goal is to determine whether a higher indicator-to-factor ratio will result in a higher rate of convergent and admissible solutions in AC research. Using a research methodology similar to that of Lance, Lambert, et al. (2004), I will reanalyze existing AC CDCE matrices taken from the available literature, reclassifying their unique constructs into three taxonomies of broad dimension factors. These categories are the 7 dimensions of Arthur et al. (2003), the 4 dimensions of Borman & Brush (1993), and the 2 dimensions of Shore et al. (1990). By reassigning the empirical dimensions into broad dimension factors, I intend to increase the indicator-to-factor ratios of the models and subsequently achieve higher rates of convergence and admissibility. As I move into progressively broader categorizations, I expect that the rate of fit will improve, as the indicator-to-factor ratios will be getting higher. Furthermore, Hoffman et al. (2011) demonstrated that these taxonomies nested into one another, which adds more support to the argument that rates of admissibility should increase with each successive categorization (i.e. a move from Arthur's seven categories to Shore's two categories for the same study will result in an even higher indicator-to-factor ratio). Given this theoretical support, I hypothesize the following:

H1: The number of convergent and admissible solutions in previously reported studies will increase as model indicator-to-factor ratio increases.

H2: Models with higher indicator-to-factor ratios will have better model fit than those with lower indicator-to-factor ratios.

H3: As indicator-to-factor ratios increase, the contribution of dimension variance will increase.

If increased indicator-to-factor ratios have an impact on dimension variance, then it is likely to find clearer and more consistent results on the impact of design characteristics on dimension variance than have been found in the past. Therefore, a reassessment of the effects of characteristics on dimension convergent validity on models with increased indicator-to-factor ratios is in order. In the same vein as Woehr & Arthur (2003), I will examine the effects of the different characteristics on construct validity in that it is predicted that same dimension, different exercise (SDDE) correlations will be higher for characteristics beneficial to dimension construct validity and different dimension, different exercise (DDDE) correlations will be lower.

H4: With fewer rated dimensions, SDDE correlations will be higher and DDDE correlations will be lower than with more rated dimensions.

H5: With lower participant-to-assessor ratios, SDDE correlations will be higher and DDDE correlations will be lower than with higher participant-to-assessor ratios.

H6: With the across-exercise rating approach, SDDE correlations will be higher and DDDE correlations will be lower than with the within-exercise rating approach.

H7: When using psychologists as raters, SDDE correlations will be higher and DDDE correlations will be lower than when using managers as raters.

H8: When training is reported, SDDE correlations will be higher and DDDE correlations will be lower than when training is not reported.

H9: When a higher number of hours of rater training is reported, SDDE correlations will be higher and DDDE correlations will be lower than when a lower number of hours is reported.

H10: When ACs function for the purpose of training/development, SDDE correlations will be higher and DDDE correlations will be lower than when ACs function for the purpose of selection/promotion.

H11: When behavioral checklists are used, SDDE correlations will be higher and DDDE correlations will be lower than when other methods of rating are used.

Table 2.1

Classification of Broad Dimensions Based on Popular Taxonomies

Arthur et al. (2003)	Borman and Brush (1993)	Shore et al. (1990)
Problem solving	Technical activities/mechanics of management	Performance style
Organizing and planning		
Drive	Other useful behavior	
Tolerance for stress and uncertainty		
Influencing others	Leadership and supervision	Interpersonal style
Communication	Communication and interpersonal facilitation	
Consideration of others		

CHAPTER 3

METHODS

Literature Review and Inclusion Criteria

A literature review was conducted using the electronic databases Academic Search Complete, Business Source Complete, Google Scholar, and PsycINFO for published and unpublished AC research, namely for those studies that include MTMM matrices. Search terms included *assessment center ratings and multitrait-multimethod, postexercise dimension rating correlations,* and *assessment center and CDCE*. An effort was made to include those papers referenced in similar studies, namely those cited in Bowler & Woehr (2006), Lance, Lambert, et al. (2004), and Lievens and Conway (2001) to ensure comprehensiveness. I contacted authors for those studies that included summary tables, but not actual CDCE matrices, as well as on the basis of abstracts of papers from Society for Industrial and Organizational Psychology conferences (2005-2012).

In order to be included in this study, the collected studies had to meet several criteria. Firstly, studies needed to report sample sizes and descriptions of all rated dimensions and exercises. Studies needed to include a minimum of three rated dimensions due to the nature of this investigation and for identification purposes. Additionally, studies had to include ratings for each dimension individually within each exercise (PEDRs), as well as correlations or covariances among the PEDRs in the form of a CDCE matrix. Excluded studies failed to meet these criteria in that they (a) provided ambiguous descriptions of rated dimensions or exercises, (b) only contained summary ratings, or (c) did not include CDCE matrices. A total of 25 studies

with 34 CDCE matrices were found to be suitable for inclusion, with 21 of these studies (84%) coming from peer-reviewed journals. The other studies are former SIOP presentations and unpublished theses and doctoral dissertations. Over 60 additional studies were initially considered for inclusion, but were deemed to be unsuitable due to the aforementioned criteria. The list of accepted studies and the number of dimensions and exercises reported can be found in Table 3.1.

Coding

Broad dimension factors. The unique dimensions used in each study were coded into three schemes of broad dimension factors: Arthur et al.'s seven factors (2003), Borman and Brush's four factors (1993), and Shore et al.'s two factors (1990). The dimensions were first recoded using Arthur et al.'s comprehensive list of AC dimensions and their respective overarching factors (2003). In the case of uncertainty, I consulted subject matter experts as to the appropriate category for the dimension in question until consensus was reached. However, this only occurred in the instance of one unique dimension and was quickly resolved by the SMEs. Using the Arthur dimensions, each dimension as then recoded into the Borman & Brush (1992) and Shore et al. (1990) taxonomies, respectively (see Table 1). All three of these taxonomies have been well-supported in the literature and have been shown to be nested within one another (Hoffman et al., 2011).

Characteristics. In line with the results of Lievens (1998) and Woehr and Arthur (2003), I examined the effects of the following characteristics on dimension and exercise variance in the retained models: (a) number of dimensions, (b) participant-assessor ratio, (c) rating approach (across-exercises or within-exercises), (d) rater occupation (psychologists vs. managers), (e) whether raters received training, (f) the length of rater training, the purpose of the AC

(administrative or developmental), and (g) whether raters used behavioral checklists. Table 3.2 shows the measurement scales for each of the characteristics.

Analyses

Using LISREL 8.80 (Jöreskog & Sörbom, 2004), I ran CFAs for each data set using their original coding schemes as well as each of the three previously described dimension taxonomies. Each model was run using maximum likelihood estimation procedures and the CDCE approach, which examines correlated dimensions and correlated exercises.

In order that models converge and reach admissible solutions, they must meet several requirements; specifically, a model that converges could still be inadmissible, indicating poor fit (Marsh, 1994). Therefore, retained models had proper solutions in that they contained no negative uniquenesses or standardized factor loadings and factor correlations greater than the absolute value of 1.00. The number of proper models for each taxonomy and model type was recorded. I ran the Rindskopf procedure (Rindskopf & Rose, 1988) which involves the addition of orthogonal dummy indicators to the model in an attempt to achieve admissibility for those models that neared admissibility (i.e. negative uniqueness values that very closely approached zero). This procedure only proved fruitful in the case of the Sagie model (the first dataset from this study, or “Sagie1”) under the Arthur et al. classification scheme.

In order to examine the effect of broad dimension factors on model acceptability, I ran multiple logistic regression using convergence and admissibility status as the dependent variable. The status was coded as a dichotomous variable, being convergent and admissible or nonconvergent and/or inadmissible. This latter category included all those models which failed to converge, were inadmissible, or both. The system of broad dimension factors served as the

categorical independent variable with four levels: original coding, Arthur et al. coding, Borman & Brush coding, and Shore et al. coding.

As a second step, I compared the fit of those models that were retained after the initial step. There is no one universal fit index to indicate which model is "best" overall. However, using a combination of fit indices can help one cover different areas of the fit spectrum and paint a fuller picture with regards to the suitability of a model (Hu & Bentler, 1999; Tanaka, 1993). Therefore, I conducted likelihood-ratio chi-square difference test to examine incremental fit, as this is the only statistic available for the assessment of model fit (Hu & Bentler, 1999). I also conducted comparisons of the Aikake information criteria (AIC) and Bayesian information criteria (BIC) in order to compare model fit. Additionally, I examined the root mean square error approximation (RMSEA), standardized root mean square residual (SRMR), Tucker-Lewis index (TLI), and comparative fit index (CFI) of each convergent and admissible model as stand-alone fit indices.

Estimation of Variance and Characteristics

In order to calculate dimension and exercise variance components, I squared dimension and exercise factor loadings. By squaring these loadings, one can avoid issues involving the addition of negative and positive factor loadings. I then converted each of the squared loadings to z-scores using Fisher's r to z transformation and found the mean of the z-scores for dimensions squared loadings and exercise squared loadings, respectively. This conversion is intended to prevent the possibility of an upward bias in mean correlation as oft occurs when taking the mean of raw correlations coefficients (Lance, Lambert, et al., 2004; Silver & Dunlap, 1987). Finally, I back-converted the z-score averages to the unstandardized *r* averages. In addition to finding the

mean-squared dimension and exercise loadings for each retained model, I found the overall mean dimension and exercise variance terms under each classification scheme.

In order to examine the effects of the various AC measurement and design characteristics on dimension construct validity, I ran a series of t-tests and correlation analyses. As a first step, I confirmed the normality of the distributions of the dimension and exercise variances under each classification scheme using the Shapiro-Wilk test. For each dichotomously measured characteristic, I ran two t-tests with the characteristic as the independent variable and proportion of variance accounted for as the dependent variable in the case of dimensions and exercises, respectively. The same was performed with each continuously measured characteristic, but instead using correlation analysis. The analytic method for each characteristic can be found in Table 3.2.

Table 3.1

Studies for Reanalysis

Study	#Dimensions	#Exercises
Arthur et al., 2000	9	3
Becker, 1990	5	4
Bowler & Woehr, 2009	13	5
Brannick et al., 1989	5	2
Brummel et al., 2009	8	6
Bycio et al., 1987	8	5
Fredericks, 1989	8	3
Haaland & Christiansen, 2002	5	5
Jackson et al., 2007	7	3
Jansen & Stoop, 2001	8	2
Joyce et al., 1994	7	4
Kleinmann, 1997	3	2

Kleinmann & Koller, 1997	3	3
Kleinmann et al., 1996	3	3
Kolk et al., 2003	3	4
Kudisch et al., 1997	6	4
Lance, Foster et al., 2004	8	3
Lance, Foster et al., 2007	6	3
Lievens et al., 2009	5	4
Parker, 1991	15	3
Pittman, 1998	4	3
Sagie & Magnezy, 1987	5	3
Schleicher et al., 2002	3	3
Schneider & Schmitt, 1992	3	4
Van der velde et al., 1994	10	3

.

Table 3.2

AC Characteristics and Proposed Method of Analysis

Characteristic	Dichotomous/Continuous	Analysis Method
Number of dimensions	Continuous	Correlation
Participant-to-assessor ratio	Continuous	Correlation
Rating approach	Dichotomous (across vs. within exercises)	t-test
Rater type	Dichotomous (psychologists vs. managers)	t-test
Rater training	Dichotomous (yes or no)	t-test
Length of training	Continuous	Correlation
AC purpose	Dichotomous (selection vs. training)	t-test
Behavioral checklists	Dichotomous (yes or no)	t-test

CHAPTER 4

RESULTS

Table 4.1 shows all convergent and admissible models under each classification scheme as well as their respective model fit indices. As predicted, the number of convergent and admissible solutions increased as the classification schemes became more broad (2 solutions using original coding[2]; 6 solutions using Arthur et al.'s coding; 7 solutions using Borman & Brush's coding; and 11 solutions using Shore et al.'s coding). Additionally, under the broadest classification scheme (Shore et al.), there were only three nonconvergent models. Despite this apparent support for Hypothesis 1, it was necessary to determine if the difference in convergence and admissibility rates was significant. Logistic regression analysis was performed in order to determine if this increase in admissible solutions was significant. The omnibus test of model fit yielded a likelihood ratio chi-square of 7.157 (df=3; p=.067; α=.10). This demonstrates that the specified model had a significant improvement in fit from the null model (containing no predictors). This indicates that there was a change in convergence and admissibility rates as indicator-to-factor ratio increased, supporting hypothesis 1. As the omnibus test was significant, the next step was to examine the individual path coefficients (Cohen, Cohen, West, & Aiken, 2012). The coefficients for Borman & Brush (B=1.429, p<.10) and Shore et al. (B=1.908, p<.05) were both significant, indicating that the convergence and admissibility rates of the tested models increased when moving from original coding to the Borman & Brush and Shore et al. coding schemes, respectively. Therefore, Hypothesis 1 was supported.

[2] As these two solutions (Arthur et al., 2000; Lievens et al., 2009) did not change when transitioning between their original dimension coding and the Arthur et al. classification scheme, no further analyses of solely original models are reported.

The next step was to compare the fit of models across dimension schemes. However, there were very few studies that were found to be admissible and even fewer studies that were retained across multiple dimension schemes (see Table 4.1). Therefore, the examination of Hypothesis 2 was limited to several studies, specifically Arthur, Kudisch, Lievens, and Sagie2 (under all three broad dimension factor schemes) and Becker and Haaland (under the Borman & Brush and Shore et al. dimension schemes). The results of this examination are presented in Tables 4.2 and 4.3.

The stand-alone fit indices (see Table 4.2) show that most models indicated acceptable fit (Kline, 2011). The mean model fit indices across the Arthur et al. dimension scheme were indicative of good model fit, while the means for the other dimension schemes indicated acceptable fit. Although mean fit became worse when moving from smaller to larger indicator-to-factor ratios, this difference in fit is not substantial. This suggests that in terms of stand-alone fit, a higher indicator-to-factor ratio did not lead to better model fit.

Overall, the results for comparative fit indicated improvement over the Arthur et al. coding, but comparable fit for the Borman & Brush and Shore et al. schemes. As provided in Table 4.3, the AIC and BIC were fairly evenly split as to the best fitting models amongst the three broad dimension factor schemes. For the AIC, two studies supported the Arthur et al. scheme, one study supported the Borman & Brush scheme, one model showed no difference between Arthur et al. and Borman & Brush, and one model showed no difference between Borman & Brush and Shore et al. The results for AIC therefore indicate that using either the Arthur et al. or Borman & Brush schemes would yield the best fit. For the BIC, one study supported Arthur et al., one supported Borman & Brush, three supported Shore et al., and one supported either Borman & Brush or Shore et al. The BIC therefore indicates that the best model

fit is achieved when using the Shore et al. dimension scheme. When using the chi-square difference test, the results were once again split, supporting both the Borman & Brush and Shore et al. classifications for best model fit. I tested each of the available schemes for each study against each other, and the best-fitting models are indicated in bold type. The best-fitting models supported the use of the Shore et al. model in three cases and either Borman & Brush or Shore et al. in three cases. The Shore et al. models were not predominantly better-fitting as to indicate that the further broadening of dimension factors substantially influenced model fit. Overall, the different comparisons across indices yielded different results, and Hypothesis 2 was not supported.

Table 4.4 contains the dimension and exercise variance components for each of the retained models under all three classification schemes. Contrary to what was predicted, dimension variance showed decreases and exercise variance, increases when moving into broader dimension factors on both the individual-study and mean levels. The Arthur study had higher dimension than exercise variance in the model using the Arthur et al. dimension factors, but showed the same pattern as all other studies under the Borman & Brush and Shore et al. classifications. The results of t-tests comparing variance components between the coding schemes (Arthur et al. vs. Borman & Brush, Arthur et al. vs. Shore et al., Borman & Brush vs. Shore et al.) indicated that the only significant change in variance was a decrease in dimension variance from the Arthur et al. vs. Shore et al. comparison, which is contrary to what was expected ($p=.043$). Hypothesis 3 not only failed to be supported, but results took the opposite form of the hypothesized effect.

The results of the t-tests and correlation analyses conducted to examine Hypotheses 4 through 11 can be found in Table 4.5. When considering the number of dimensions, there were

two significant results: for dimension variance under the Arthur et al. factor scheme ($r=0.809$, $p<.05$) and for exercise variance under the Borman & Brush scheme ($r=-0.839$, $p<.05$). The first significant correlation went in the opposite direction of the hypothesized effect, which would suggest that dimension variance becomes higher when *more* dimensions are rated, rather than less. The second significant correlation shows that exercise variance became higher when less dimensions were rated. This effect was hypothesized to have the reverse implication for exercise variance. Although all of the other correlations were negative, as hypothesized, none of them were significant, failing to support Hypothesis 4.

When considering participant-to-assessor ratio, results went in the expected direction, with dimension variance being higher than exercise variance under each of the dimension schemes. However, none of these correlations were significant, providing no support for Hypothesis 5.

The t-tests on the categorically measured characteristics yielded similarly disappointing results. The examination of rating approach yielded no significant results under any classification scheme, nor did that of rater type. Therefore, Hypotheses 6 and 7 were not supported.

In the case of rater training, analyses could not be conducted using the Arthur et al. broad dimension factors, as all admissible solutions came from studies that reported the administration of training. None of the results for the other two classification schemes were significant; therefore Hypothesis 8 was not supported. Furthermore, none of the correlation coefficients between dimension or exercise variance and length of rater training were significant, and all of the dimension variance correlations were negative, contrary to the hypothesized effect. Hypothesis 9 was not supported.

I obtained a significant t-test result when considering AC purpose for dimension variance under the Borman & Brush scheme, $t(4)=18.852$, $p<.05$. As hypothesized, this suggests that those ACs operating for developmental and training purposes result in higher dimension variance than would administrative ACs. However, as none of the results under the other coding schemes were significant, Hypothesis 10 does not receive overall support.

It was possible to analyze the effect of behavioral checklists on dimension and exercise variance under the Arthur et al. classification scheme, as it contained the only admissible study that used behavioral checklists (Jackson et al., 2007). However, the results of the t-test were not significant and did not support Hypothesis 11. Overall, none of the predicted hypotheses could be supported.

In addition to the previous analyses, I ran sign tests to determine if the overall direction of the results was in the expected direction (Pagano, 2004). The results can be found in Table 4.6. The expected direction for each of the variance components refers to the predictions made in Hypotheses 4 through 11. The obtained direction refers to either (a) the direction of the estimated correlation between the characteristic and the coding scheme or (b) the facet of the characteristic in question with the highest average variance component. A "+" indicates that the data supported the expected direction, whereas a "-" indicates that the data went in a different direction than was expected. For example, it was expected that ACs that utilized across-exercise dimension ratings would yield higher dimension variance than those ACs that used other rating approaches. However, the average exercise variance component for the within-exercise approach was higher than those of the across-exercise or combined approaches. Accordingly, this result was coded with a "-".

These tests did not yield significant results according to the binomial distribution (Pagano, 2004). Both dimension and exercise variance achieved the same number of results in the expected direction. Accordingly, no definitive conclusion could be drawn regarding the directionality of the results. The next section contains possible rationale for the expected findings.

Table 4.1

Convergence and Admissibility of Tested Models

	Arthur et al.			Borman & Brush			Shore et al.		
	C & A	C & I	NC & IA	C & A	C & I	NC & IA	C & A	C & I	NC & IA
Arthur	x			x			x		
Becker			x	x			x		
Bowler		x				x		x	
Brannick 1		x			x			x	
Brannick 2		x			x			x	
Brummel 1		x			x		x		
Brummel 2		x				x		x	
Bycio		x				x		x	
Fredricks		x				x		x	
Haaland		x		x			x		
Jackson	x				x			x	
Jansen		x			x		x		
Joyce		x				x	x		
1Klein97			x			x		x	
2Klein97			x			x		x	
KleinKoller		x			x			x	
KleinKuptsch1			x			x			x
KleinKuptsch2			x			x			x
Kolk 1		x			x			x	
Kolk 2			x			x	x		
Kudisch	x			x			x		
1Lance04		x			x			x	
2Lance04		x			x			x	
Lance07		x			x			x	
Lievens	x			x			x		
Parker		x		x				x	
Pittman		x				x			x
Sagie 1	x				x			x	
Sagie 2	x			x			x		
Schleicher 1			x			x		x	
Schleicher 2			x			x		x	
Schneider		x			x			x	
Van der velde			x		x			x	

Note. C=convergent; A=admissible; NC=nonconvergent; IA=inadmissible

Table 4.2

Model Fit Indices--Stand-Alone

	Study	chi-sq	df	RMSEA	CONF. INT.	SRMR	TLI	CFI	AIC	BIC
Arthur et al.	Arthur	13.600	33	0.000	{0.000,0.000}	0.019	1.020	0.990	103.240	238.778
	Jackson	78.720	66	0.030	{0.000,0.053}	0.034	1.000	0.980	186.270	366.947
	Kudisch	215.58**	121	0.062	{0.043,0.079}	0.066	0.910	0.940	322.030	555.561
	Lievens	120.61**	51	0.043	{0.033,0.053}	0.029	0.970	0.980	229.970	478.31
	Sagie1	129.36**	69	0.047	{0.032,0.060}	0.035	0.970	0.970	221.220	426.033
	Sagie2	157.56**	69	0.056	{0.044,0.068}	0.040	0.960	0.960	251.220	459.697
Borman & Brush	Arthur	17.21	36	0.000	{0.000,0.000}	0.022	1.020	1.000	101.190	227.376
	Becker	227.76**	95	0.081	{0.065,0.098}	0.079	0.890	0.920	312.390	522.841
	Haaland	255.54**	155	0.077	{0.054,0.098}	0.089	0.880	0.910	379.140	587.618
	Kudisch	227.82**	124	0.069	{0.052,0.086}	0.071	0.910	0.930	337.110	553.019
	Lievens	154.09**	54	0.049	{0.040,0.059}	0.029	0.950	0.970	255.090	491.917
	Parker	814.19**	366	0.054	{0.049,0.060}	0.054	0.930	0.940	974.550	1402.01
	Sagie2	157.56**	69	0.056	{0.044,0.068}	0.040	0.960	0.980	251.220	459.697
Shore et al.	Arthur	46.75	38	0.047	{0.000,0.079}	0.036	0.990	0.990	130.450	246.908
	Becker	227.76**	95	0.081	{0.065,0.098}	0.079	0.890	0.920	312.390	522.841
	Brummel1	554.42**	359	0.046	{0.027,0.061}	0.062	0.950	0.960	644.880	1042.57
	Haaland	250.68**	157	0.066	{0.040,0.088}	0.082	0.890	0.920	358.510	574.019
	Jansen	113.38	81	0.023	{0.010,0.033}	0.028	0.990	0.990	220.350	472.014
	Joyce	328**	200	0.074	{0.053,0.094}	0.084	0.870	0.900	433.130	656.129
	Kolk2	53.69*	35	0.065	{0.000,0.100}	0.064	0.920	0.960	135.710	251.28
	Kudisch	234.59**	126	0.071	{0.054,0.087}	0.072	0.900	0.930	340.280	549.934
	Lievens	158.52**	56	0.049	{0.040,0.058}	0.030	0.950	0.970	256.240	483.099
	Sagie2	163.65**	71	0.057	{0.045,0.069}	0.042	0.960	0.980	254.680	371.121

Note.Chi-sq=chi-squared statistic; df=degrees of freedom; conf. inf.=RMSEA confidence interval; *=p<.05; **=p<.01

Table
4.3

Mean Variance Components

		Dimension Var.	Exercise Var.
Arthur et al.	Arthur	0.510	0.207
	Jackson	0.194	0.604
	Kudisch	0.215	0.421
	Lievens	0.158	0.453
	Sagie1	0.216	0.403
	Sagie2	0.137	0.414
	M	0.238	0.417
Borman & Brush	Arthur	0.306	0.424
	Becker	0.115	0.476
	Haaland	0.136	0.511
	Kudisch	0.205	0.423
	Lievens	0.100	0.455
	Parker	0.053	0.331
	Sagie2	0.136	0.415
	M	0.150	0.434
Shore et al.	Arthur	0.257	0.422
	Becker	0.115	0.476
	Brummel1	0.039	0.646
	Haaland	0.140	0.456
	Jansen	0.050	0.474
	Joyce	0.208	0.342
	Kolk2	0.093	0.616
	Kudisch	0.173	0.430
	Lievens	0.076	0.468
	Parker	0.030	0.332
	Sagie2	0.129	0.409
	M	0.119	0.461

Note. M=total mean

Table 4.4

Comparative Model Fit--Best Models

	Models	AIC	BIC	Comparison	chi-diff	df diff	sig.	chi-best
Arthur	A, B, S	**B**	**B**	A vs. B	3.61	3	0.31	A/B
				B vs. S	29.54	2	0.00	S
				A vs. S	33.15	5	0.00	**S**
Becker	B, S	**B/S**	**B/S**	B vs. S	0.00	0	x	**B/S**
Haaland	B, S	**S**	**S**	B vs. S	4.86	2	0.09	**B/S**
Kudisch	A, B, S	**A**	**S**	A vs. B	12.24	3	0.01	B
				B vs. S	6.77	2	0.03	**S**
Lievens	A, B, S	**A**	**A**	A vs. B	33.48	3	0.00	B
				B vs. S	4.43	2	0.11	**B/S**
Sagie 2	A, B, S	**A/B**	**S**	A vs. B	0.00	0	x	A/B
				B vs. S	6.09	2	0.05	**S**

Note. A=Arthur et al.; B=Borman & Brush; S=Shore et al.; **Bold** denotes best-fitting model; chi-diff=chi-square difference test; df diff=difference in

degrees of freedom; sig.=significance; chi-best=best-fitting model from chi-sqare difference test

Table 4.5

Correlations & t-tests of AC Design Characteristics and Variance Components

	Arthur et al.		Borman & Brush		Shore et al.	
	Dimension	Exercise	Dimension	Exercise	Dimension	Exercise
No. dimensions	*.890**	*-0.454*	*-0.157*	*-0.839**	*-0.217*	*-0.458*
P-to-A ratio	*0.989*	*-0.993*	*0.41*	*-0.192*	*0.558*	*-0.503*
Rating approach	0.539	1.417	0.545	0.115	0.910	0.020
Rater type	0.227	2.291	2.757	1.186	1.199	0.245
Rater training	n/a	n/a	0.029	2.836	0.058	0.014
Length of training	*-0.732*	*0.847*	*-0.558*	*-0.709*	*-0.426*	*-0.136*
AC purpose	0.956	0.209	18.852*	4.462	1.177	1.329
Behavioral checklists	0.103	4.346	n/a	n/a	n/a	n/a

Note. *=p<.05; *italics* indicate correlation analysis; plain text indicates t-test; n/a indicates could not be estimated

Table 4.6

Sign Test Results

	Dimension Variance: Obtained	Dimension Variance: Expected	**Dimension Variance: Difference**	Exercise Variance: Obtained	Exercise Variance: Expected	**Exercise Variance: Difference**
# of dimensions	Negative correlation	Negative correlation	+	Negative correlation	Negative correlation	+
Participant-assessor ratio	Positive correlation	Negative correlation	-	Negative correlation	Negative correlation	+
Rating approach	Across-exercises	Within-exercises	-	Within-exercises	Within-exercises	+
Rater type	Psychologists and managers	Psychologists	-	Psychologists and managers	Psychologists	-
Training	Training reported	Training reported	+	Training reported	Training reported	+
Length of training	Negative correlation	Positive correlation	-	Negative correlation	Positive correlation	-
AC purpose	Developmental	Developmental	+	Developmental	Developmental	+
N=7			***p*(3+)=0.2743**			***p*(5+)=0.1641**

Note. + indicates that variance component went in the expected direction; - indicates that variance component did not favor the expected direction; *p*=probability in the binomial distribution=.50.

CHAPTER 5

DISCUSSION

For years, AC researchers have been plagued with difficulty in establishing construct validity for dimensions (Arthur et al., 2003; Bycio et al., 1987; Cahoon et al., 2012; Lance et al., 2000). This challenge stems, in part, from a dearth of convergent and admissible model solutions when using the theoretically suitable CDCE model structure (Lance, Lambert et al., 2004). It has recently been suggested that the construct validity problem could stem from the methodological artifact of specifying too few indicators per factor in models (Hoffman et al., 2011; Monahan, 2011; Tomás et al., 2000).

This study reexamined previously published models under the assumption that higher indicator-to-factor ratios would positively influence convergence and admissibility rates. This ratio increase was accomplished through the reassignment of originally reported dimensions into broad dimension factors—broad categories of dimensions that are representative of a wide number of unique dimensions. Due to similarities in unique dimensions measured in ACs, reclassification into broad dimension factors allows the unique dimensions to serve as indicators of broader dimension factors, increasing the indicator-to-factor ratio (Hoffman et al., 2011).

This study demonstrated a significant effect of increased indicator-to-factor ratios (and broad dimension factors) on the subsequent convergence and admissibility rates of models. However, for those models that were convergent and admissible with higher indicator-to-factor ratios, exercise variance still dominated dimension variance, providing no remedy to the

construct validity problem in ACs. Therefore, although this study attempted to contribute to the literature by demonstrating the prevalence of dimension variance when using measurement and design modifications, it showed that exercise variance is still dominant. If research continues to demonstrate the importance of exercise variance, it should follow that administrators revise measurement and reporting procedures to reflect that the effects we are seeing are those of the exercises, not of dimensions. In other words, ACs should be based around tasks in their structure, measurement, and feedback mechanisms (Jackson et al., 2007; Lance, 2008).

Limitations

There were a number of limitations in this study that could have contributed to its results. The main limitation was sample size—at all levels of analysis. To begin with, there is not an abundance of MTMM and AC studies, severely limiting the pool of available models I had to work with at the onset of this process. Additionally, the unexpectedly low number of retained models resulted in difficulty when analyzing the role of design characteristics. Since it was anticipated that the higher indicator-to-factor ratios would have a very strong impact on nonconvergence and inadmissibility, examination of design characteristics on the retained models seemed feasible. In order for this section of the study to have had conclusive and clear results, it would have been beneficial for more models to have been retained as admissible. When examining Hypotheses 4 through 12, the limited sample size proved to be a hindrance in examining the effects of characteristics on AC construct validity. Various characteristics, such as use of behavioral checklists and AC purpose, proved to be identical for most or all studies under certain classifications, making it impossible to examine the effect of the characteristics. Therefore, the limited number of studies available in the first place was a challenge to this study,

and the further dwindling of availability as the study progressed made achieving significant results even more difficult.

On a related note, the obtained results likely suffered from problems with model specification. Although the use of broad dimension factors increased each model's indicator-to-factor ratio, it is likely, if not certain, that this increase was not substantial enough in certain models. A recent MTMM study using both real and simulated data found that the commonly used ratio of three indicators-per-factor failed to reach convergent and admissible solutions over 80% of the time (Dumenci & Yates, 2012). However, these problems were eradicated when increasing the ratio to five indicators-per-factor. In the present study, an increase in indicator-to-factor ratio did not necessarily guarantee even three indicators per factor, even when using the Shore et al. dimension scheme, due to the original AC design employed in each of the studies. Therefore, my difficulty in reaching the predicted results was probably at least partially attributable to the use of too few indicators per factor, despite the use of broad dimension factors.

In addition, the complexity behind the design and administration of ACs may have served as a limitation to this study. As there are no standard operating procedures for ACs, each center runs with its own nuances and idiosyncrasies (Howard, 2008). Therefore, there is likely a plethora variables at play beyond the design characteristics examined in this study that impacts AC ratings. If a move is made to standardize AC design and measurement across the field, it is possible that researchers will come to more consistent results in their analyses of construct validity and other AC-related phenomena. For example, Monahan (2011) achieved results supporting the use of higher indicator-to-factor ratios, but used the same AC for all measurements in the study, meaning that the differences at play when using various ACs were

not relevant. As this was not the case in this study, the differences between ACs could have been harmful to the eventual results.

Implications and Future Directions

Despite a dearth of significant findings, this study demonstrated that using higher indicator-to-factor ratios may affect model fit to a point—the Borman & Brush and Shore et al. models were found to have comparable fit, but still generally had better fit than the Arthur models when considering comparative fit indices. Additionally, a significant pattern of increased convergence and admissibility was established when using higher indicator-to-factor ratios Specifically, under the Shore et al. dimension scheme, which was the broadest factor scheme used, there were the most admissible models and the fewest nonconvergent models. These results do not help to eradicate the construct validity paradox, but they do offer a bit of hope to the process of finding a solution.

Furthermore, AC designers should consider the implications of constructing ACs so that subsequent models have at least five indicators per factor (Dumenci & Yates, 2012). There were too few models in the current study that adhered to this guideline, which could have accounted for the low convergence and admissibility rates, even after increasing model indicator-to-factor ratios. It is possible that if ACs are designed to have a minimum of five indicators-per-factor, an effort such as the current study could yield more definitive results.

Another under-explored area of research is the use of hybrid models in ACs (Hoffman, 2012; Hoffman et al., 2011). A recently proposed hybrid model contains exercises, broad dimension factors, and a general performance factor, with exercises and general performance serving as uncorrelated first-order factors (Hoffman et al., 2011). This model was found to be better fitting than the CDCE model and without nonconvergence and inadmissibility issues.

Additionally, it is still theoretically relevant to AC research. Although dimension variance was still dominated by exercise variance, this hybrid model and other proposed models should be further explored in order to combat nonconvergence and inadmissibility problems.

When considering the results of both this study and the existing body of AC construct validity literature, the best course of action may finally be to move away from dimensions and start focusing on AC exercises. This might not have been the original intention for the use of ACs, but exercises have repeatedly demonstrated their worth in the AC context—they continue to dominate dimensions in construct validity, despite numerous interventions intended to remedy such results. Perhaps ACs do not suffer from a construct validity "problem", but rather a misdirection in the search for construct validity (Lance, 2008b).

This study is the most recent in a long line of studies attempting to salvage AC dimensions (e.g. Bowler & Woehr, 2006; Hoffman et al., 2011; Lievens & Conway, 2001) without achieving its desired results. Despite treatment by some researchers of exercises as error (Lievens & Conway, 2001), it is clear that exercises play an important, if not leading, role in ACs. It may be time for practitioners to come to terms with the idea that dimension effects are simply not as robust as exercise effects, regardless of methodological and design considerations.

Task-based ACs (TBACs) have been suggested since the 1980s as a solution to the AC construct validity problem (Jackson, 2012; Lance, 2008). TBACs have assessors rate participants on task lists within each exercise, instead of focusing on personality-related dimensions within exercises (Jackson, 2012). The behavioral ratings on the task lists are combined in order to form exercise scores, which serve as samples of role-relevant behavior. TBACs suggest criterion-related validity with supervisory ratings of employee performance and are theoretically representative of what AC research has been demonstrating for years (Lance, 2012). However,

researchers have shied away from TBACs, likely due to their nontraditional approach to AC ratings. Accordingly, more criterion-related validity evidence of TBACs needs to be established, as well as more research as to TBACs' general properties and best practices in comparison to traditional dimension-based ACs (DBACs; Lance, 2012). Although DBACs are widely accepted and trusted by researchers, practitioners, and participants, the overwhelming evidence that has been presented over the years of the cross-situational specificity in ACs has been consistently ignored. Therefore, further examination of TBACs could only serve to improve upon a much-valued administrative and developmental tool (Jackson, 2012; Jackson et al., 2007; Lance, 2012).

Conclusion

Although the use of higher indicator-to-factor ratios demonstrated a significant pattern of improvement on model nonconvergence and inadmissibility in ACs, dimension variance was still dominated by exercise variance. Instead of achieving its goal of demonstrating the prevalence of dimension variance, this study only further acknowledged the dominating role of AC exercises in construct validity. Accordingly, it may be time to step away from AC dimensions and continue working toward the improvement of TBACs.

REFERENCES

Anderson, N., Lievens, F., Van Dam, K., & Born, M. (2006). A construct-driven investigation of gender differences in a leadership-role assessment center. *Journal Of Applied Psychology, 91,* 555-566. doi: 10.1037/0021-9010.91.3.555

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

Arthur, W. R., Day, E., & 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. doi:10.1111/j.1754-9434.2007.00019.x

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

Bagozzi, R. P., & Yi, Y. (1991). Multitrait-multimethod matrices in consumer research. *Journal Of Consumer Research, 17,* 426-439. doi: 10.1086/208568

Bandalos, D. L., & Boehm-Kaufman, M. R. (2009). Four common misconceptions in exploratory factor analysis. In C. E. Lance, R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 61-87). New York, NY: Routledge/Taylor &

Francis Group.

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

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

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. doi:10.1037/0021-9010.91.5.1114

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

Brannick, M. T., Michaels, C. E., & Baker, D. P. (1989). Construct validity of in-basket scores. *Journal Of Applied Psychology, 74,* 957-963. doi:10.1037/0021-9010.74.6.957

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

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

Cahoon, M. V., Bowler, M. C., & Bowler, J. L. (2012). A reevaluation of assessment center construct-related validity. *International Journal of Business and Management, 7,* 3-19.

doi: 10.5539/ijbm.v7n9p3

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81-105. doi:10.1037/h0046016

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

Cohen, J., Cohen, P., West, S., & Aiken, L. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. New York: Routledge.

Dumenci, L. & Yates, P.D. (2012). Nonconvergence/improper solution problems with the correlated-trait correlated-method parameterization of a multitrait-multimethod matrix. *Educational and Psychological Measurement, 72,* 800-807. doi: 10.1177/0013164412443540

Fredricks, A. J. (1989). Assessment center ratings: models and process (Unpublished doctoral dissertation). University of Nebraska—Lincoln, Lincoln, NE.

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. doi:10.1037/0021-9010.72.3.493

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. doi:10.1037/0021-9010.74.4.611

Guenole, N., Chernyshenko, O. S., Stark, S., Cockerill, T., & Drasgow, F. (2013). More than a mirage: A large-scale assessment centre with more dimension variance than exercise

variance. *Journal Of Occupational & Organizational Psychology*, *86*, 5-21. doi:10.1111/j.2044-8325.2012.02063.x

Haaland, S., & Christianse, N. D. (2002). Implications of trait-activation theory for evaluating the construct validity of assessment center ratings. *Personnel Psychology, 55*, 137-163. doi: 10.1111/j.1744-6570.2002.tb00106.x

Hoffman, B.J. (2012). Exercises, dimensions, and the great battle of Lilliput: Evidence for a mixed model interpretation of AC performance. In D.R. Jackson, C.E. Lance, B.J. Hoffman (Eds.), *The Psychology of Assessment Centers* (pp. 281-306). New York: Routledge.

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

House, R. J., Filley, A. C., & Kerr, S. (1971). Relation of leader consideration and initiating structure to R and D subordinates' satisfaction. *Administrative Science Quarterly, 16*, 19-30.

Howard, A. (2008). Making assessment centers work the way they are supposed to. *Industrial And Organizational Psychology: Perspectives On Science And Practice*, 98-104. doi:10.1111/j.1754-9434.2007.00018.x

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55. doi:10.1080/10705519909540118

Jackson, D.R. (2012). Task-based assessment centers: Theoretical perspectives. In D.R. Jackson, C.E. Lance, B.J. Hoffman (Eds.), *The Psychology of Assessment Centers* (pp.

173-189). New York: Routledge.

Jackson, D. R., Barney, A. R., Stillman, J. A., & Kirkley, W. (2007). When traits are behaviors: The relationship between behavioral responses and trait-based overall assessment center ratings. *Human Performance, 20,* 415-432.

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

Jöreskog, K., & Sörbom, D. (2004). LISREL 8.70. Chicago, IL: Scientific Software International Inc.

Joyce, L. W., Thayer, P. W., & Pond III, S. B. (1994). Managerial functions: An alternative to traditional assessment center dimensions? *Personnel Psychology, 47,* 109-121. doi: 10.1111/j.1744-6570.1994.tb02412.x

Kleinmann, M. (1997). Transparenz der anforderungsdimensionen: Ein koderator der Konstrukt- und kriteriumsvalidität des assessment-centers. *Zeitschrift für Arbeits-und Organisationspsychologie, 41,* 171-181.

Kleinmann, M., & Köller, O. (1997). Construct validity of assessment centers: Appropriate use of confirmatory factor analysis and suitable construction principles. *Journal Of Social Behavior & Personality,* 12(5), 65-84.

Kleinmann, M., Kuptsch, C., & Köller, O. (1996). Transparency: A necessary requirement for the construct validity of assessment centres. *Applied Psychology: An International Review, 45,* 67-84. doi:10.1111/j.1464-0597.1996.tb00849.x

Klimoski, R., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology, 40,* 243-260. doi:10.1111/j.1744-

6570.1987.tb00603.x

Kline, R.B. (2011). *Principles and practices of structural equation modeling.* New York: Guilford Press.

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

Kudisch, J. D., Ladd, R. T., & Dobbins, G. H. (1997). New evidence on the construct validity of diagnostic assessment centers: The findings may not be so troubling after all. *Journal of Social Behavior & Personality, 12*(5), 129-144.

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

Lance, C.E. (2012). Research into task-based assessment centers. In D.R. Jackson, C.E. Lance, B.J. Hoffman (Eds.), *The Psychology of Assessment Centers* (pp. 218-233). New York: Routledge.

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. doi:10.1037/0021-9010.89.1.22

Lance, C., Foster, R., Nemeth, Y., Gentry, W., & Drollinger, S. (2007). Extending the nomological network of assessment center construct validity: prediction of cross-situationally consistent and specific aspects of assessment center performance. *Human Performance, 20*, 345-362.

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised

estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal Of Applied Psychology, 89,* 377-385. doi:10.1037/0021-9010.89.2.377

Lance, C. E., 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. doi:10.1207/S15327043HUP1304_1

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

Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal Of Selection And Assessment, 6*(3), 141-152. doi:10.1111/1468-2389.00085

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. doi:10.1037/0021-9010.86.6.1202

Lievens, F., Dilchert, S., & Ones, D. S. (2009). The importance of exercise and dimension factors in assessment centers: Simultaneous examinations of construct-related and criterion-related validity. *Human Performance, 22,* 375-390. doi:10.1080/08959280903248310

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling, 1,* 5-34. doi: 10.1080/10705519409539960

Marsh, H. W., Hau, K., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number

of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181-220. doi:10.1207/s15327906mbr3302_1

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (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. doi:10.1037/0021-9010.93.5.1042

Monahan, E.L. (2011). The search for the mythical assessment center dimension: Measurement artifact vs. substantive conclusion (Unpublished master's thesis). University of Georgia, Athens, GA.

Neidig, R. D., & Neidig, P. J. (1984). Multiple assessment center exercises and job relatedness. *Journal Of Applied Psychology, 69,* 182-186. doi:10.1037/0021-9010.69.1.182

Pagano, R. R. (2004). *Understanding Statistics in the Behavioral Sciences*. Belmont: Wadsworth.

Parker, M. L. W. (1991). A construct validation of the Florida Principal Competencies Assessment Center using confirmatory factor analysis (Unpublished doctoral dissertation). University of South Florida: Tampa, FL.

Pittman, S. (1998). An examination of construct validity within an assessment center (Unpublished master's thesis). George Mason University, Fairfax, VA.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23,* 51-67. doi:10.1207/s15327906mbr2301_3

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. doi:10.1207/S15327043HUP1304_2

Russell, C. J. (1985). Individual decision processes in an assessment center. *Journal Of Applied Psychology, 70*, 737-746. doi:10.1037/0021-9010.70.4.737

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

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. doi: 10.1111/j.2044-8325.1997.tb00634.x

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal Of Applied Psychology, 87*, 735-746. doi:10.1037/0021-9010.87.4.735

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. doi:10.1037/0021-9010.77.1.32

Shore, T. H., Thornton III, G. C., & McFarlane Shore, L. (1990). Construct validity of two categories of assessment center dimension ratings. *Personnel Psychology, 43*, 101-116. doi: 10.1111/j.1744-6570.1990.tb02008.x

Silver, N., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used?. *Journal Of Applied Psychology, 72*, 146-148. doi:10.1037/0021-9010.72.1.146

Spychalski, A. C., Quiñones, M. A., Gaugler, B. B., & Pohley, K. (1997). A survey of assessment center practices in organization in the United States. *Personnel Psychology*,

50, 71-90. doi: 10.1111/j.1744-6570.1997.tb00901.x

Tanaka, J.S. (1993). Multifaceted conceptions of fit in structural equation models. In K.A. Bollen & J.S. Long (Eds.), *Testing Structural Equation Models* (pp. 10-39). Newbury Park, CA: Sage.

Thornton, G.C., III, & Byham, W.C. (1982). *Assessment centers and managerial performance.* New York: Academic.

Thornton, G., & Gibbons, A. M. (2009). Validity of assessment centers for personnel selection. *Human Resource Management Review*, *19*, 169-187. doi:10.1016/j.hrmr.2009.02.002

Tomás, J.M., Hontangas, P.M., & Oliver, A. (2000). Linear confirmatory factor models to evaluate multitrait-multimethod matrices: The effects of number of indicators and correlation among methods. *Multivariate Behavioral Research, 35,* 469-499. doi: 10.1207/S15327906MBR3504

Van der Velde, E. G., M. P. Born, & Hofkes, K. (1994). Begripsvalidering van een assessment center met behulp van confirmatieve factoranalyse. [Construct validity of an assessment center using confirmatory factor analysis.]. *Gedrag en Organisatie, 7*, 18-26.

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*, 231-251. doi:10.1037/1082-989X.3.2.231

Woehr, D. J., & Arthur, W. R. (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. doi:10.1177/014920630302900206