THE SEARCH FOR THE MYTHICAL ASSESSMENT CENTER DIMENSION: MEASUREMENT ARTIFACT VS. SUBSTANTIVE CONCLUSION

by

ELIZABETH L. MONAHAN

(Under the Direction of Charles E. Lance)

ABSTRACT

This study investigated the addition of multiple indicators to solve the convergence and admissibility issues associated with models of the internal structure of Assessment Centers. Specifically, we used behavioral checklist items rather than traditional Post Exercise Dimensions Ratings as manifest indicators of multi-trait multi-method analysis. These behavioral checklist items were used to create two separate types of ratings: Behavioral Checklist Sum ratings and Disaggregated Behavioral Checklist ratings. Behavioral Checklist Sum ratings are composed of the sum of all of the ratings of each dimension within each exercise. Disaggregated Behavioral Checklist ratings are mini-composites formed for each exercise dimension combination. However, neither the model of the traditional Post Exercise Dimension ratings nor the model that used Behavioral Checklist Sum ratings as manifest indicators converged to admissibility. These results suggest that the frequent failure to find dimensions in models of the internal structure of Assessment Centers is merely a methodological artifact.

INDEX WORDS: Assessment Center, MTMM, CTCM, Construct Validity

THE SEARCH FOR THE MYTHICAL ASSESSMENT CENTER DIMENSION: MEASUREMENT ARTIFACT VS. SUBSTANTIVE CONCLUSION

by

ELIZABETH L. MONAHAN

B. S., University of Georgia, 2009

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

2011

THE SEARCH FOR THE MYTHICAL ASSESSMENT CENTER DIMENSION:

MEASUREMENT ARTIFACT VS. SUBSTANTIVE CONCLUSION

by

ELIZABETH L. MONAHAN

Major Professor:	Charles E. Lance
Committee:	Brian J. Hoffman Karl W. Kuhnert

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2011

DEDICATION

To my parents, who instilled in me the importance of an education at a young age. And finally, to my family and best friends, who provided me with the appropriate encouragement and support, to which I contribute my success.

ACKNOWLEDGEMENTS

First, I would like to thank both Brian Hoffman and Charles Lance for providing support and constructive feedback for this project. Additionally, I would like to collectively acknowledge my committee members, Charles Lance, Brian Hoffman, and Karl Kuhnert, without whose guidance, this project would not have materialized. Finally, I would like to acknowledge Mark Foster, who provided me with access to data, without which this project could not have been conducted.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

CHAPTER 1

INTRODUCTION

For over forty years, Assessment Centers (ACs) have been used for selection and development (Thornton, 1992). The typical AC format aims for a candidate to be evaluated on behaviors relating to certain dimensions assessed in several exercises. Those participants are then given post exercise dimension ratings (PEDRs), which are ratings made on performance of a dimension within an exercise. These ratings are often used in AC validity research.

The content validity of ACs is evidenced through job-relevant exercises (Klimoski & Brickner, 1987; Thornton & Byham, 1982). Additionally, ACs routinely display criterion-related validity (Arthur, Day, McNelly, & Edens, 2003; Gaugler, Rosenthal, Thronton, & Bentson, 1987; Hoffman, Melchers, Messal, Kleinmann, & Ladd, in press; Meriac, Hoffman, Woehr, & Fleisher, 2008; Sackett, 1987; Sackett & Tuzinski, 2001). However, based on commonly used analyses, limited evidence exists for the construct validity of ACs.

ACs have historically been constructed to measure specific performance dimensions using multiple exercises. Typically, AC researchers use the multitrait-multimethod (MTMM; Campbell & Fiske, 1959) framework to investigate the construct validity of AC ratings for analyses. Using the MTMM framework, it is anticipated that analysis of AC PEDRs should reveal high correlations of the same dimension across different exercises (SDDE; convergent validity) and low correlations for different dimensions in the same exercise (DDSE; discriminant validity). But, most often, the PEDRs measure more of the effect of the exercise than that of the dimension (Bycio, Alvares, & Hahn, 1987; Sackett & Dreher, 1982). This is demonstrated by larger DDSE correlations than SDDE correlations, leading some to conclude that ACs lack construct validity (Bycio et al., 1987; Chan, 1996; Sackett & Dreher, 1982; Schneider & Schmitt, 1992; Turnage & Muchinsky, 1982).

AC researchers have generally failed to produce an admissible solution for the correlated trait – correlated method (CTCM) model that contains all exercises and all dimensions. The current study examines an alternative to PEDRs that may allow for the convergence and admissibility of the CTCM model for AC dimension ratings. If a convergent and admissible solution of the CTCM model is found, those AC researchers who support the importance of dimensions will have a stronger argument for their retention. Given that ACs are designed around dimensions, dimensions are key for feedback, and AC research routinely uses dimensions to represent AC performance, it is critical to direct research at explanations for lack of support for dimensions. Support for dimensions, and specifically for multiple dimensions, has been shown independent of the typical MTMM validity based searches (Hoffman et al., in press; Meriac et al., 2008). Therefore, this study aims to clarify whether the lack of success of the CTCM model to fit AC data is due to (a) the fact that the full CTCM model is inappropriate, or (b) a methodological artifact – the typical CTCM parameterization has too few indicators. Determining the source of the lack of success will help solve an often debated AC issue. If the full CTCM model proves inappropriate, researchers will have evidence that ACs do not fit the multiple dimension framework provided in the CTCM model. On the other hand, if this study provides an admissible solution of the CTCM model, the “construct validity paradox” (Arthur, Woehr, & Maldegan; 2000) often mentioned in the literature, is in fact, merely a derivative of a methodological artifact. Therefore, at a time when many are doubting the importance of dimensions (Jackson, Barney, Stillman, & Kirkley, 2007; Lance, Newbolt, Gatewood, Foster, French, & Smith, 2000), this study seeks to provide evidence for multiple dimensions by providing a convergent and admissible CTCM model. In doing so, this study aims to present a model that emphasizes the importance of both exercises and dimensions.

CHAPTER 2

LITERATURE REVIEW AND HYPOTHESES

MTMM, CTCM, and Confirmatory Analysis of ACs

For years, ACs have been designed to elicit consistent dimensions from performance across exercises. At the end of each exercise, PEDRs are made based on perceived performance on a particular dimension within a particular exercise. Due to the apparent similarities of the AC structure with the MTMM framework, much validity research has focused on the application of MTMM based analyses to AC data in the form of PEDRs. In fact, Campbell and Fiske's (1959) seminal work has become the foundation of modern day thinking about construct validity. Using the MTMM framework as applied to ACs, "dimensions serve as traits and exercises as methods" (Sackett & Dreher, 1982, p. 402). Since the Campbell and Fiske (1959) criteria are subjective, other approaches have been considered, such as path analysis (Avison, 1978; Schmitt, 1978).

However, the most common application to the traditional MTMM method (Campbell & Fiske, 1959) is confirmatory factor analysis (CFA; Marsh, 1989; Tomás, Hontangas, & Oliver, 2000; Widaman, 1985, 1992). This is most likely due to the notion that CFA allows the researcher to better analyze MTMM matrices by overcoming the criteria limitations of the Campbell and Fiske (1959) approach (Kalleberg & Kluegel, 1975). The CFA model also allows the total variance to be separated into method, trait, and error components (Jöreskog, 1974). Furthermore, CFA allows the researcher to model correlations among factors and fix factor loadings consistent with theoretical rationale (Kleinmann & Köller, 1997).

Although a variety of CFA based MTMM models have been specified, the CTCM model has emerged as the most theoretically appropriate for ACs. In fact, Lance and colleagues (2002) argued that this model is the closest model conceptually to the Campbell and Fiske (1959) matrix. This model has been shown to be superior to alternate models in terms of the accuracy and magnitude of the factor loadings. In the CTCM model as analyzed by CFA, all of the method factors and all of the trait factors are allowed to correlate among themselves, but correlations between trait factors and method factors are set to zero. However, the CTCM model suffers known problems, the most damning of which are, (a) identification and estimation problems, (b) unstable solutions (unless extremely large samples are utilized), and (c) the confounding of the trait and method variance caused by the increase of correlations between traits and between methods (Bagozzi, 1993; Brannick & Spector, 1990; Kenny & Kashy, 1992; Kumar & Dillon, 1992; Lance, Noble, & Scullen, 2002). The chance of these problems occurring increases when data are missing, when there are high correlations between measured variables, when sample size is small, when the model is misspecified, when variables can load on more than one latent factor, or when latent factors have a small number of indicators (Marsh, 1989). It is noteworthy that the conditions under which the general CTCM model suffers from convergence and admissibility issues correspond closely to the typical AC MTMM data. Given the failure of this model to return proper solutions when investigating attitudes, personality, and performance measurement, it is not surprising that ACs are also plagued with improper solutions (Hoffman, in press).

The frequency with which the CTCM model results in improper solutions has led many researchers to use models such as the correlated trait-correlated uniquenesses (CTCU) model (Kenny 1976, 1979; Lance et al, 2002; Marsh, 1989). Using the CTCU model, method factor correlations are restricted to zero. Inflation of the covariance between traits and trait variance is inevitable (Byrne & Goffin, 1993; Kenny & Kashy, 1992; Lance et al., 2000) because the correlated method effects surface as general trait variance (Marsh & Bailey, 1991). Due to the faults of the CTCU model, mainly the incorrect assignment of variance among its components, it is necessary to explore additional solutions. Given that the CTCM model is often deemed the most theoretically appropriate model, research to improve the admissibility of this model in particular is warranted.

Graham and Collins (1991) suggested that the addition of redundant trait-method indicators may solve the convergence and identification problems of the CTCM model. Adding redundant indicators, while maintaining the same number of factors, changes the indicator-to-factor ratio. Research using CFA of MTMM matrices with varying indicator-to-factor ratios has found that when this ratio is high, more accurate parameter estimates are obtained, along with fewer nonconvergent solutions and improper solutions (Marsh, Hau, Balla, & Grayson, 1998). Including more indicators per factor increases the degrees of freedom allowing for a more powerful test than a model with fewer indicators (MacCallum, Browne, Sugawara, 1996). This increase of degrees of freedom allowed Tomás and colleagues (2000) to conclude that when two or more indicators per trait-method combination were used in the CTCM model of MTMM matrices there was a decrease in the number of ill-defined solutions. Therefore, a potential solution to avoiding improper solutions is increasing the indicator-to-factor ratio.

In the AC field, a proper solution of the CTCM model will have many benefits. Not only will a proper solution in itself be a unique contribution, but it will also allow for the comparison of other models that are often found as the “best-fitting” models. In fact, most often, the model chosen in internal structure analyses as the best model is agreed upon simply due to the lack of admissible models with which to compare that model. Providing a solution with both exercises and dimensions represented, will give AC researchers the ability to determine which model is truly most representative of ACs.

Dimensions vs. Exercises

Support for AC performance dimensions has been found (Arthur et al., 2003; Hoffman et al., in press; Meriac et al., 2008); however, very little evidence exists for dimensions in internal structure analyses. In fact, AC researchers seem unsatisfied with evidence for dimension variance, while reports of significantly large exercise variance are far too common (Bycio et al. 1987; Lance et al, 2000; Sackett & Dreher, 1982). As such, it is often the case that exercises are seen as the only important component in ACs (Jackson et al., 2007; Lance et al., 2000).

One of the arguments against the importance of dimensions is that of situational specificity. The cross-situational *specificity* argument (Lance et al., 2000; Neidig & Neidig, 1984; Turnage & Muchinsky, 1982) holds that different situational characteristics cause variations in an individual's behavior, whereas cross-situational *consistency* is the idea that an individual's behavior is stable across situations and exercises serve as alternative ways to measure the same traits. Unlike method bias, which captures unwanted variance, the situational specificity interpretation contends that exercise effects provide "situationally specific, performance-relevant variance" (Lance et al., 2000, p. 327). The cross-situational consistency argument, on the other hand, is the reason AC researchers utilize MTMM based analyses.

For years, AC researchers have used MTMM based internal structure analyses to unsuccessfully provide sufficient construct validity evidence. Even more disconcerting is that AC researchers have used incorrect models and misguided statistical manipulations to arrive at different conclusions about dimensions and exercises. Lievens and Conway's (2001) review used a CTCU model to show support for dimensions by revealing equal proportions of variance accounted for by both exercises and dimensions (34%). Lance et al. (2004) later reanalyzed these data using the CTCM model which frequently resulted in improper solutions. Lance et al. (2004) adapted this model by removing the dimensions and adding a general performance factor. Note that this model (one trait and correlated methods, 1TCM) essentially increases the ratio of indicators to factors, a practice that, as previously discussed, has been shown to be effective in increasing convergence and admissibility (Marsh et al., 1998). The model containing the correlated exercises and general performance factor revealed 52% of variance to be explained by exercises and 14% to be explained by the general factor. Lance et al. (2004) blamed the unique results of Lievens and Conway (2001) on the utilization of the CTCU model which is known for overestimating dimension variance (Kenny & Kashy, 1992; Marsh & Bailey, 1991). Since both of the aforementioned studies (Lance et al. 2004; Lievens et al. 2001) used individually analyzed MTMM matrices, Bowler and Woehr (2006) meta-analyzed these MTMM matrices into a single matrix. CFA results of this matrix (using a modified CTCM model) revealed the best model to be that of all exercises and all dimensions with 22% of variance explained by dimensions and 33% of variance explained by exercises. However, these results were only ascertained through the "modification" (fixing latent factor intercorrelations to zero) of both the CTCM model and the 1TCM, which consequently permitted convergence. Often reaching in their techniques, AC researchers have continually attempted to circumvent the convergence problems of the CTCM models.

In light of these convergence problems, Lance, Woehr, and Meade (2007) used Monte Carlo generated AC data to be representative of exercise and dimension effects. However, the

CFA-based MTMM results did not support the CTCM model. In fact, due to convergence and admissibility issues, other models were deemed to better fit the data. Interestingly, the problems concerning the convergence of the CTCM models are not specific to AC research (e.g., Bagozzi, 1993). In fact, Lance, Dawson, Birkelbach, and Hoffman (2010) deliberately avoided using AC data and found convergence for only 30% of the CFA-based MTMM models tested. Consequently, it should not be assumed that the existence of dimensions in ACs is inhibiting the convergence and proper identification of the CTCM model being that solution admissibility and convergence are not necessarily appropriate criteria for judging the correct population model (Lance et al., 2007).

Another reason thought to cause the undesirably large exercise effects is that AC researchers do not apply systematic psychometric evaluation to the espoused constructs (Arthur & Villado, 2008). In fact, more often that not, dimensions are accepted as being reflective of the label assigned (Brannick, 2008). It is important not only theoretically but practically to ensure that the dimensions being measured are actually what AC designers are attempting to measure (i.e., establishing construct validity). If the dimensions are not in fact what they were intended to be, incorrect hiring decisions could be made and suggested developmental improvements may be inappropriate (Howard, 2008). As stated by Arthur, Day, & Woehr (2008; p. 106), “The fundamental issue here is one of construct validity and an emphasis on the fact that merely labeling data as reflecting a particular construct (espoused construct) does not mean that is the construct that is being assessed (actual construct).” Up to this point, AC researchers and practitioners have neglected to psychometrically treat AC dimensions as other tested constructs and confirm that they measure the necessary components of the job.

Recently, Hoffman and colleagues (in press) proposed a hybrid model (made of broad dimensions, general performance, and exercises) which combined similar dimensions into broader dimensions to circumvent the notion that AC dimensions lack discriminant validity. This model intended to decrease correlations among dimensions through the collapsing of similar dimensions and through increasing the number of manifest indicators for each dimension factor. All of the proposed models converged and were admissible. The current study utilized multiple indicators to determine if the success of the hybrid model (Hoffman, et al., in press) was in fact due to the inclusion of the broad dimension factors, or if the success can be replicated with solely the addition of more manifest indicators.

Behavioral Checklists as Indicators

Unfortunately, researchers are in a conundrum when it comes to increasing the number of indicators. In many ACs, PEDRs are the lowest unit of analysis, prohibiting the derivation of additional manifest indicators. A more recent AC phenomenon is the use of behavioral checklists (Donahue, Truxillo, Cornwell, & Gerrity, 1997; Reilly, Henry, & Smither, 1990). On behavioral checklists, important behaviors are listed under each dimension for each exercise. Behavioral checklists aim to increase the dimension construct validity and reduce the possible cognitive loads that assessors may experience due to the large amount of observations they are expected to rate. If the behaviors necessary for each dimension are written out in checklist form, the raters do not have to use the definition of the dimension to place behaviors in categories. This can be helpful in reducing the cognitive load, thereby increasing the accuracy of observations (Donahue et al., 1997; Reilly et al., 1990). Using behavioral checklists has been shown to increase convergent validity (Reilly et al., 1990), discriminant validity (Donahue et al. 1997; Reilly et al., 1990), and interrater reliability (Hennessy, Mabey, & Warr, 1998).

Although the intention of using behavioral checklists is to reduce cognitive demands on raters and increase rating accuracy (Donahue et al. 1997; Reilly et al. 1990), they may also be useful as manifest indicators to aid in the elimination of convergence and admissibility problems of the CTCM model. The checklist items can be used instead of the PEDRs as input to analyses. There are multiple ways these checklist items can be used to form ratings. Behavioral checklist sum (BCS; Donahue et al., 1997) ratings are formed from summing the results of the individual behavioral checklists and averaging those sums across raters. Using this method, the raters are not discussing disagreements between each other and their ratings are unaltered unlike in the PEDRs. The PEDRs and BCS ratings both have one indicator per exercise-dimension combination. Although these ratings are formed differently, it is expected, since they both contain one indicator per exercise-dimension combination, that both sets of ratings (PEDRs and BCS) will result in comparable results. Another manipulation utilizes the items of the behavioral checklist to form multiple indicators called disaggregated behavioral checklist (DBC) ratings. Using the DBC ratings, different results can be anticipated since these ratings provide multiple manifest indicators for each exercise-dimension combination.

The PEDRs utilized in the present study came directly from the AC reported. These ratings provide one manifest indicator for each exercise and dimension combination. BCS ratings are formed by summing the checklist items for each dimension within each exercise. These ratings also provide one manifest indicator for each exercise dimension combination. DBC ratings are formed from mini-composites of the checklist items, but provide multiple manifest indicators for each exercise-dimension combination.

H1: Analyses of the MTMM matrix of PEDRs will reveal that the most appropriate model is the model of three exercises and one dimension (Model 2; 1DCE).

H2: Analyses of the MTMM matrix composed of BCS ratings will reveal the most appropriate model is the model of three exercises and one dimension. (Model 2; 1DCE).

H3: Analyses of the MTMM matrix of composed of DBC ratings will allow for convergence and admissibility of all models and will reveal the best fit for the model of three methods exercises and six dimensions. (Model 1; 6D3E).

CHAPTER 3

METHOD

Sample

This study used archival AC ratings for 254 law enforcement officers of corporal rank. These officerrs participated in an AC in either 2007 or 2008 that was used for promotion decision purposes. Assessors were law enforcement officers (from a different state) at least one rank higher than the participant. Raters received eight to ten hours of general rater training which included: definition of dimensions they would rate, how to use behavioral checklists, and the appropriate way to take notes and observe correctly. Raters used behavioral checklists in each exercise for each of the dimensions measured. The training also included a frame-of-reference training component which helps increase the accuracy of ratings (Woehr & Huffcutt, 1994). Groups of raters were assigned to a certain exercise: one group rated dimensions in the first exercise, a second group rated the next exercise, and a third group rated the last exercise.

AC Procedure

The exercises in this AC were constructed to give a setting for important job-related dimensions. Each exercise was designed to provide attendees with tasks that they seem as relevant to the job. The AC consisted of the following K=3 exercises: oral presentation, role play, and written exercise. In the oral presentation exercise, participants were asked to deliver a plan to resolve given problems. In the role play exercise, participants were asked standardized questions by a simulated supervisor. In the written exercise, participants were asked to summarize the credentials that would make them a commendable leader in their organization. The dimensions used in this AC were intended to provide insight of individual performance beyond just good or bad performance in an exercise. The AC consisted of the following J=6 dimensions (defined in the Appendix): perception, decisiveness, judgment, oral communication, leadership, organizational planning, and written communication.

Behavioral checklists were used for each dimension within each exercise. Each assessor was instructed to rate behaviors on a scale from one to seven. At the end of every exercise, each assessor individually assigned AC participants a rating for every dimension in that exercise. The three raters must then collectively reach consensus on these PEDRs.

Disaggregation (Multiple Indicators, Parcels)

The disaggregation of items is necessary to provide more indicators to allow for the convergence and proper identification of the correlated dimension correlated exercise (CDCE, i.e., CTCM) model. Although the individual checklist items could be used as manifest indicators, the behavioral checklists all have varying numbers of items. Since using individual items does not necessarily provide better solutions than parcels composed of items (Marsh et al., 1998), the current study utilized parcels as indicators. The use of parcels allowed for each exercise-dimension combination to have the same number of manifest indicators.

When several indicators of a given construct are summed and averaged, a parcel is formed (Cattell, 1956). From the list of items on the scale, without reusing individual items, multiple parcels are made until all items are assigned to a parcel. These parcels may then be used as indicators, in a first order factor to represent the latent construct of interest (Kishton & Widaman, 1994). The popularity of item-parcels has increased and the use of item-parcels has been documented in the areas of mental ability (Widaman, Gibbs, & Geary, 1987), personality (Marsh & Gouvernet, 1989), and measurement (Velicer, Huckel, & Hansen, 1989).

Marsh et al (1998) found more reliable factors, fewer nonconvergent solutions, reduced item idiosyncrasies, more accurate parameter estimates, fewer improper solutions, and greater interpretability when more indicators per factor were used. They also found that the solutions achieved when using item parcels as indicators were not better than the solutions achieved when using the individual items as indicators. Therefore, item parceling may circumvent the complications (i.e. improper solutions, nonconvergent solutions, etc.) and give similar information to using individual items.

The DBC ratings were formed in a manner similar to the "Partial Disaggregation Model" of Bagozzi and Edwards (1998). To create the DBC ratings, the individual items on each behavioral checklist were combined such that each dimension in each exercise had four indicators. First, each item on the checklist was summed and averaged across the three raters, so that in essence, a single checklist was analyzed. Randomly, the items on the behavioral checklist were assigned into four groups (parcels). The utilization of four indicators (parcels) is due to the fact that some dimensions had behavioral checklists with only four items. Consequently, there are some instances where a single item acts as an indicator. Once there were four indicators for each dimension, two degree of freedom CFAs were performed. It was expected that for each dimension in each exercise, the four created parcels loaded on the intended dimension.

Analyses

In order to compare the resulting fits of the models of different ratings, the following analyses were performed for each type (PEDRs, BCS ratings, and DBC ratings) of rating. The first operationalization used the traditional PEDRs as manifest indicators. The rationale for the use of the PEDRs was strictly to serve as a baseline and to ensure that traditional modeling yielded an improper solution in this same as it has in past research. The second operationalization used the BCS (total of the behavioral checklist items) as manifest indicators. The rationale for the use of BCS ratings is to ensure that the success of the DBC ratings is due to the increased number of indicators, rather than the use of behavioral checklist items. The final operationalization used the item parcels formed (DBC ratings) based on the behavioral checklist as manifest indicators. To test whether convergence and admissibility are in fact due to the addition of more indicators and not simply the use of checklist information, the analyses of the BCS ratings were compared with the analyses of the DBC ratings.

The correlations among the six dimensions and three exercises were input into LISREL 8.7 (Jöreskog & Sörbom, 2004). In order to assess convergent and discriminant validity for dimensions, a number of different CFA models were fit. This practice is consistent with past research (e.g. Bycio et al., 1987; Donahue et al., 1997; Lance et al., 2000). To fit six correlated dimension factors and three correlated exercise factors, Model 1 (K-exercise-J-dimension model) was created. Model 1 reflects the common view that AC ratings are representative of both the dimension and exercise. To fit three correlated exercises and one general performance factor, Model 2 (K-exercise-1-dimension model) was created. Comparing Model 2 and Model 1 serves as a test of discriminant validity of AC dimensions. The assumption for Model 2 is that AC raters are assessing performance as an overall measure, therefore not differentiating among multiple dimensions. To fit three correlated exercises and no dimensions, Model 3 (K-exercise-0-dimension model) was created. Comparison of Model 3 with Model 1 provides an omnibus test of convergent validity for AC dimensions. Model 3 assumes raters are not cognizant of performance consistencies across exercises. To fit one general exercise factor and six correlated dimension factors, Model 4 (1-exercise-J-dimension model) was created. Comparison of Model 1 and Model 4 tests whether exercises are truly different from one another (i.e. discriminability of the exercise factors). To fit zero exercises and six correlated dimension factors, Model 5 (0-exercise-J-dimension model) was created. Comparing Model 5 against Model 1 provides an omnibus test of the presence of exercise effects. The five models tested are summarized in Table 4.1.

LISREL 8.7's (Jöreskog & Sörbom, 2004) completely standardized maximum likelihood parameter estimates were examined to determine if solutions were proper. Those models that were proper were evaluated according to (a) the comparative fit index (CFI; Bentler, 1990) (b) the chi-square statistic (c) the root mean squared error of approximation (RMSEA; Steiger, 1990) (d) standardized root mean squared residual (SRMR; Bentler, 1995) and (e) Bentler and Bonett's (1980) nonnormed fit index (NNFI), which is the Tucker-Lewis index (NNFI/TLI, Tucker & Lewis, 1973). To test convergent validity, discriminant validity, and for the presence of exercise effects, difference chi-square ($\Delta\chi^2$) tests were conducted.

Additionally, I sought to psychometrically test the representativeness of the checklist items of their intended dimension. To do so, CFAs were conducted on each behavioral checklist. Just as one would expect an item of a test to load on a particular construct, the parcels created for each dimension in each exercise were expected to load on the intended dimension. These indicator-dimension relationships will be described based on their significance loadings and model fit.

CHAPTER 4

RESULTS

Table 4.2 shows CFA model fit indices for all ratings. As predicted (Hypothesis 1 and Hypothesis 2) the models containing PEDRs and BCS ratings yielded similar results. In fact, both sets of ratings had models that produced improper solutions. For the models composed of PEDRs, those models that contained six correlated dimension factors were either nonconvergent or inadmissible. More specifically, Model 1 (K-exercise-J-dimension model) was nonconvergent and Model 4 (1-exercise-J-dimension model) and Model 5 (0-exercise-J-dimension model) converged, but were inadmissible. For the models composed of BCS ratings, all models containing six correlated dimension factors (Model 1, Model 4, and Model 5) were nonconvergent. Consequently, all J-dimension models were considered to be inconsistent with the data. Since a key requirement in the evaluation of model fit is that the model results in a proper solution (see Marsh, 1994), these models were eliminated as possibly fitting the data. Because only two models were proper (for both models of PEDRs and BCS ratings), tests of discriminant validity could not be conducted. Of the two models that were proper, the model that provided the best fit to the data, according to commonly reported overall goodness-of-fit indices (Hu & Bentler, 1998, 1999), was Model 2 (K-exercise-1-dimension model). These results are consistent with previous research (e.g. Lance et al., 2000; Lance et al., 2004). Therefore, both Hypothesis 1 and Hypothesis 2 were supported.

All of the models that were created using DBC ratings converged to admissible solutions. Therefore, it can be assumed that the multiple manifest indicators allowed for the proper solutions of all models tested. As proposed (Hypothesis 3), the best fitting model was Model 1

(K-exercise-J-dimension). The standardized factor loadings of the model are shown in Table 4.5. Since all models converged to proper solutions, chi-square difference tests were conducted to determine convergent validity of dimensions, discriminant validity of dimensions, and to test for the presence of exercise effects. A chi-square difference test between Model 1, which specified three exercises and six dimensions, and Model 2, which specified three exercises and one dimension, can be regarded as an omnibus test of discriminant validity of dimensions in the AC. The difference, $\Delta\chi^2_{15}$= 796.24, p<.05, supported discriminant validity of dimensions and supported Model 1 as the better fitting model. A chi-square difference test between Model 1 and Model 3, which specified only three exercises, can be regarded as an omnibus test of convergent validity of dimensions in the AC. The difference, $\Delta\chi^2_{60}$= 2747.63, p<.05, supported convergent validity of dimensions and proved Model 1 as the better fitting model. A chi-square difference test between Model 1 and Model 4, which specified one exercise and six dimensions, can be regarded as an omnibus test of discriminability of exercise factors. The difference, $\Delta\chi^2_{3}$= 2887.59, p<.05, supported discriminability of exercise factors and proved Model 1 to be the better fitting model. A chi-square difference test between Model 1 and Model 5, which specified only six dimension factors, can be regarded as an omnibus test of the presence of exercise effects. The difference, $\Delta\ \chi^2_{66}$= 150301.81, p<.05, supported the presence of exercise effects and proved Model 1 to be the better fitting model.

Indicator-dimension fit for the DBC indicators is shown in Table 4.3 and factor loadings of the indicators are provided in Table 4.4. All of the parcels significantly loaded on the intended dimension. Additionally, twelve out of the fifteen CFAs conducted for indicator-dimension fit were acceptable according to commonly used standards (Hu & Bentler, 1998, 1999).

Interestingly, the worst fitting models (leadership measured in role play, judgment measured in oral presentation, and organizational planning measured in oral presentation), all had behavioral checklists that had four items (did not require parceling). This difference may be explained by the fact that the use of parcels reduced measurement error associated with a given indicator. Therefore, indicators composed of more than one item had improved fit.

The standardized factor loadings from the 6D3E model of the DBC ratings are shown in Table 4.5. It is worth mentioning that the exercise factor loadings were generally much larger than dimension effects. Past research has revealed similar results for models that use PEDRs as input (Bycio et al., 1987; Lance et al., 2000; Lance et al., 2004). However, it was never anticipated that the addition of more indicators would change the magnitude of effects, just that they would allow for a convergent and admissible solution.

Even more interesting is that it appears that each dimension is primarily measured in a single exercise. For instance, judgment, decisiveness, and leadership appeared to be predominantly measured in the oral presentation exercise, while perception, organizational planning, and oral communication were mostly measured in the role play exercise.

The latent factor correlations of the 6D3E model for the DBC ratings are shown in Table 4.6. The strongest correlations among dimensions are between decisiveness and judgment, and judgment and leadership. Only two exercises (written exercise and oral presentation) were moderately correlated. The remaining exercise intercorrelations were small. Some dimensions in particular were highly correlated. In order to test for their discriminant validity, beyond the omnibus test of discriminant validity, certain constraints were made. To determine if Decisiveness and Judgment were in fact different dimensions, the model constraining this correlation to 1.0 was compared against the model in which this relationship was not constrained. The same procedure was also performed for the relationships between Leadership and Judgment and also for Leadership and Decisiveness. The results of the chi-square difference test, indicated that Decisiveness and Judgment were in fact empirically discriminable ($\Delta\chi^2_1$= 18.83, *p*<.01). Additionally, the Leadership and Decisiveness dimensions were also shown to be separate dimensions ($\Delta\chi^2_1$= 83.34, *p*<.01). The Leadership and Judgment dimensions were found to be most similar (as they were not significantly different at the .01 level), but still were found to be differentiable ($\Delta\chi^2_1$= 4.91, *p*<.05).

To determine the magnitude of variance explained by dimensions, exercises, and errors, I first transformed the standardized loadings using Fisher's r to z transformation. Once converted to z scores, the values were averaged and then back transformed. In the 6D3E model, exercises accounted for 64% of variance, dimensions accounted for 13% of variance, and error explained 10% of the variance.

Table 4.1
The Explanation of CFA Models

Model	Exercises	Dimensions	Measuring
1 (6D3E)	3 correlated exercises	6 correlated dimensions	Baseline - comparison
2 (1D3E)	3 correlated exercises	1 general dimension	Discriminant Validity
3 (0D3E)	3 correlated exercises	No dimensions	Convergent Validity
4 (6D1E)	1 general exercise factor	6 correlated dimensions	Discriminability of Exercises
5 (6D0E)	No exercises	6 correlated dimensions	Exercise Effects

Table 4.2

Models of Ratings

	Proper?	df	χ^2	SRMSR	RMSEA	TLI	CFI	$\Delta\chi^2$ vs. Model 1
Post Exercise Dimension Ratings								
Model 1[a] (6D3E)	No	57	133.96*	.038	.071	.97	.99	
Model 2 (1D3E)	Yes	72	169.10*	.035	.070	.97	.98	
Model 3 (0D3E)	Yes	87	374.24*	.12	.061	.93	.94	
Model 4[b] (6D1E)	No	60	564.72*	.10	.17	.83	.90	
Model 5[b] (6D0E)	No	75	2196.22*	.28	.38	.43	.59	
Behavioral Checklist Sum								
Model 1[a] (6D3E)	No	57	135.07*	.060	.072	.97	.99	
Model 2 (1D3E)	Yes	72	249.26*	.034	.10	.95	.97	
Model 3 (0D3E)	Yes	87	588.06*	.068	.16	.89	.91	
Model 4[a] (6D1E)	No	60	871.43*	.11	.21	.74	.85	
Model 5[a] (6D0E)	No	75	2817.08*	.31	.40	.31	.51	
Disaggregated Behavioral Checklist								
Model 1 (6D3E)	Yes	1632	5408.04*	.11	.10	.94	.95	
Model 2 (1D3E)	Yes	1647	6204.28*	.064	.13	.93	.94	796.24**
Model 3 (0D3E)	Yes	1707	8155.67*	.097	.17	.91	.91	2747.63**
Model 4 (6D1E)	Yes	1635	8295.63*	.11	.16	.90	.91	2887.59**
Model 5 (6D0E)	Yes	1698	155709.85*	.30	.32	.80	.81	150301.81**

Note. [a]Model did not converge; [b]Model converged to an inadmissible solution; * p < .00001, **p < .05. Model 1 = 3 correlated exercises 6 correlated dimensions, Model 2 = 3 correlated exercises 1 general dimension, Model 3 = 3 correlated exercises no dimensions, Model D = 4 general exercise factor and 6 correlated dimensions, Model 5 = no exercises 6 correlated dimensions

Table 4.3

Indicator-Dimension Fit

	df	χ^2	SRMSR	RMSEA	TLI	CFI
Oral Presentation						
Perception	2	6.32*	.027	.090	.97	.99
Judgment	2	78.30**	.063	.39	.69	.90
Organizational Planning	2	65.94**	.064	.34	.77	.92
Decisiveness	2	14.25**	.0095	.16	.97	.99
Oral Communication	2	27.30**	.022	.23	.92	.97
Leadership	2	8.91*	.029	.11	.98	.99
Role Play						
Perception	2	12.85**	.019	.15	.96	.99
Judgment	2	6.18*	.014	.093	.99	1.00
Organizational Planning	2	1.23	.0032	.00	1.00	1.00
Decisiveness	2	21.13**	.013	.20	.95	.98
Oral Communication	2	22.71**	.020	.19	.94	.98
Leadership	2	132.33**	.077	.45	.59	.86
Written Exercise						
Perception	2	9.68**	.026	.12	.96	.99
Judgment	2	1.69	.014	.00	1.00	1.00
Organizational Planning	2	10.12**	.034	.12	.96	.99

Note. * *p* < .05, ** *p* < .01.

Table 4.4
Factor Loadings of Parcels on Indicators

Exercise	PER	JUD	ORG	DEC	ORC	LED
Oral Presentation						
Indicator 1	.92**	.81**	.70**	.95**	.95**	.92**
Indicator 2	.78**	.74**	.93**	.93**	.90**	.77**
Indicator 3	.70**	.93**	.79**	.96**	.92**	.92**
Indicator 4	.38*	.83**	.99**	.90**	.83**	.95**
Role Play						
Indicator 1	.78**	.97**	.92**	.95**	.91**	.66**
Indicator 2	.86**	.83**	.96**	.91**	.93**	.97**
Indicator 3	.90**	.90**	.88**	.97**	.93**	1.00**
Indicator 4	.89**	.80**	.97**	.85**	.80**	.86**
Written Exercise						
Indicator 1	.79**	.71**	.52*			
Indicator 2	.69**	.76**	.95**			
Indicator 3	.92**	.95**	.92**			
Indicator 4	.76**	.56*	.72**			

Note.,PER = perception, JUD = judgment, ORG = organizational planning, DEC = decisiveness, ORC = oral communication, LED = leadership; * *p* < .05, ** *p* < .01.
.001.

Table 4.5

Standardized Parameter Estimates

DBC	PER	JUD	ORG	DEC	ORC	LED	OP	RP	WE
OP_PER									
Indicator 1	.06						.85**		
Indicator 2	.04						.74**		
Indicator 3	.15**						.69**		
Indicator 4	.20**						.41**		
OP_JUD									
Indicator 1		.27**					.84**		
Indicator 2		.15**					.89**		
Indicator 3		.62**					.73**		
Indicator 4		.62**					.65**		
OP_ORG									
Indicator 1			-.04				.80**		
Indicator 2			.19**				.86**		
Indicator 3			.05				.85**		
Indicator 4			.11**				.91**		
OP_DEC									
Indicator 1				.61**			.74**		
Indicator 2				.46**			.84**		
Indicator 3				.62**			.75**		
Indicator 4				.59**			.73**		
OP_ORC									
Indicator 1					.32**		.79**		
Indicator 2					.27**		.74**		
Indicator 3					.27**		.85**		
Indicator 4					.22**		.81**		
OP_LED									
Indicator 1						.58**	.74**		
Indicator 2						.61**	.59**		
Indicator 3						.51**	.81**		
Indicator 4						.62**	.72**		

Note. DBC= disaggregated behavioral checklist, PER = perception, OP = oral presentation, JUD = judgment, ORG = organizational planning, DEC = decisiveness, ORC = oral communication, LED = leadership, RP = role play, WE = written exercise. ** p < .01

Table 4.5 (continued)

DBC	PER	JUD	ORG	DEC	ORC	LED	OP	RP	WE
RP_PER									
Indicator 1	.46**							.61**	
Indicator 2	.55**							.66**	
Indicator 3	.59**							.66**	
Indicator 4	.55**							.68**	
RP_JUD									
Indicator 1		.09**						.96**	
Indicator 2		.04						.82**	
Indicator 3		.04						.89**	
Indicator 4		.05						.80**	
RP_ORG									
Indicator 1			.23**					.91**	
Indicator 2			.50**					.81**	
Indicator 3			.43**					.76**	
Indicator 4			.46**					.85**	
RP_DEC									
Indicator 1				.14**				.94**	
Indicator 2				.12**				.89**	
Indicator 3				.09**				.95**	
Indicator 4				.00				.86**	
RP_ORC									
Indicator 1					.68**			.57**	
Indicator 2					.69**			.60**	
Indicator 3					.67**			.64**	
Indicator 4					.61**			.50**	
RP_LED									
Indicator 1						.14**		.83**	
Indicator 2						.02		.86**	
Indicator 3						.06		.89**	
Indicator 4						.10**		.94**	

Note. DBC= disaggregated behavioral checklist, RP = role play, PER = perception, JUD = judgment, ORG = organizational planning , DEC = decisiveness, ORC = oral communication, LED = leadership, OP = oral presentation, WE = written exercise. ** *p* < .01

Table 4.5 (continued)

DBC	PER	JUD	ORG	DEC	ORC	LED	OP	RP	WE
WE_PER									
Indicator 1	.04								.80**
Indicator 2	-.03								.75**
Indicator 3	.07								.84**
Indicator 4	.11**								.71**
WE_JUD									
Indicator 1		.06							.72**
Indicator 2		-.13**							.84**
Indicator 3		.11**							.88**
Indicator 4		.03							.52**
WE_ORG									
Indicator 1			-.21**						.48**
Indicator 2			-.06						.86**
Indicator 3			-.09**						.89**
Indicator 4			-.08						.68**

Note. DBC= disaggregated behavioral checklist, WE = written exercise, PER = perception, JUD = judgment, ORG = organizational planning, DEC = decisiveness, ORC = oral communication, LED = leadership, OP = oral presentation, RP = role play, . ** $p < .01$

Table 4.6
Latent Factor Correlations

	1	2	3	4	5	6	7	8	9
1. Perception	1								
2. Judgment	.32**	1							
3. Organizational Planning	.37**	.42**	1						
4. Decisiveness	.29**	.94**	.35**	1					
5. Oral Communication	.10	.07	.59**	.08	1				
6. Leadership	.33**	.96**	.37**	.89**	.12	1			
7. Oral Presentation	.00	.00	.00	.00	.00	.00	1		
8. Role Play	.00	.00	.00	.00	.00	.00	-.04	1	
9. Written Exercise	.00	.00	.00	.00	.00	.00	.28**	.15**	1

Note. Values of .00 and 1 were all fixed. All other values were free.. ** *p* < .01.

CHAPTER 5

DISCUSSION

Most AC research to date reports convergence and admissibility problems for the CTCM model (Lance et al., 2000; Lance et al., 2004). However, research has shown that this issue is not strictly an AC problem (Lance et al., 2010). The present study reevaluated the conclusions of prior AC research by determining if problems such as lack of convergence and admissibility that arise when applying CFA MTMM models to PEDR data are attributable to (a) the fact that the full CTCM model is inappropriate or (b) a methodological artifact – the typical CTCM parameterization just has too few indicators. In contrast to prior analytic work that has failed to provide a convergent and admissible solution for the CTCM model, these results reveal a proper solution for the CTCM model. By achieving a convergent and admissible solution for the CTCM model, this study contributes to the literature by providing researchers a solution that emphasizes the importance for both exercises and dimensions. Additionally, this study suggests that the convergence issues of the CTCM model are, in fact, a methodological artifact.

Due to the apparent physical similarities of the exercise-dimension correlation structure and the MTMM matrix, researchers have applied MTMM analyses to AC ratings for years (Bycio et al., 1987; Lance et al., 2000; Sackett & Dreher, 1982; Turnage & Muchinsky, 1982). Therefore, researchers in search for construct validity have attempted to find the typical convergent and discriminant validity evidence first described by Campbell and Fiske (1959). Despite the lack of success, researchers have continued to use factor analytic derivations of the MTMM framework. Even though CFA is an improvement of the traditional MTMM framework (Campbell & Fiske, 1959), there are still flaws. Campbell and Fiske (1959) stated that when using MTMM in test validation research, a few assumptions are required: traits and methods are uncorrelated, measures are reliable, and methods are assumed to assess traits equally well (Campbell & Fiske 1959; Schmitt & Stults, 1986). But, as this study shows, often certain dimensions are captured best in a single exercise. For example, perception in role play exhibited factor loadings on average four times the size of perception loadings from the oral presentation. Even more discrepant, perception in role play exhibited factor loadings that were on average eleven times the size of perception loadings from the written exercise. This finding is consistent with the idea that AC exercises should be providing unique information about dimensions (Howard, 2008). Instead of finding a new way to determine construct validity, it seems that some researchers want to change an already-working assessment tool. To defend this working tool, we sought to prove the commonly-cited problems of applying MTMM-based analyses to AC data were not the fault of the AC, rather a methodological artifact. Accordingly, considering the overall functioning of the typical AC (Bowler & Woehr, 2009), we recommend researchers amend the analyses applied to the AC ratings rather than amend the AC itself.

By testing models that included multiple manifest indicators, this study offers an approach that is very distinct from the typical internal structure studies of ACs. Specifically, the use of multiple manifest indicators allowed not only for the convergence and admissibility of all models tested, but revealed that the CTCM model best fit the data. This finding provides evidence that the problems often encountered with the CTCM model are due to the lack of manifest indicators in past research rather than some failure of AC researchers to attain a theoretically plausible AC framework.

Recently, Hoffman and colleagues (in press) achieved a convergent, admissible solution in four independent samples. Their model collapsed manifest dimensions into broader dimension factors. Although Hoffman et al. (in press) hypothesized that this model reached a proper solution by more naturally modeling the broad dimension structure of managerial performance, our results suggest that their findings may have occurred simply by increasing the indicator-to factor-ratio. Since our study sought to provide clarification on what in particular allowed for the proper solution of a CTCM model, only the indicator-to-factor ratio was increased. Given the proper solutions achieved in this study resulted from increasing the number of manifest indicators, we can conclude that the issue of nonconvergence is grounded in the lack of the appropriate number of indicators. Therefore, the ability for the MTMM analyses of AC data to produce proper solutions is dependent on the empirical identification allotted for by the use of multiple manifest indicators.

More often than not, typical factor analytic research of CTCM models produces nonconvergent results. Additionally, other models tested are often either nonconvergent or inadmissible. Of the convergent and admissible models, the best fitting model is most often the model of one general dimensions and all exercises (1TCM). It is this common finding that has led many to believe that dimensions are not important (Crawley, Pinder, & Herriot, 1990; Jackson et al., 2007; Lance et al., 2000). As such, some researchers aim to change the structure of the traditional AC by removing dimensions completely. It is this thinking that has led to the development of task-based ACs (Jackson et al., 2005; Jackson et al., 2007; Lowry, 1996). In the task-based AC approach, the assessor rates general performance in the exercise, as opposed to rating dimensions within each exercise (Lowry, 1996; Jackson et al. 2005); thus, eliminating the behavioral inferences an assessor could make in a dimension-based AC. Given that task-based ACs have received mixed reviews, it seems important to maintain the traditional AC structure. As is evidenced through successful criterion-related and content validity searches, the traditional AC structure works. What is illustrated in this study is that the lack of construct validity is due to the underidentification of the CTCM model, rather than the failure of AC dimensions.

Like much other AC construct validity research, the variance explained by exercises greatly outweighed the variance explained by dimensions. More specifically, the variance explained by exercises and dimensions in this study was very similar to that found by Lance and colleagues (2004) who used a model with a general dimension and correlated exercises. However, it was never the intention of this study to increase the variance explained by dimensions, or decrease the variance explained by exercises. In fact, it was assumed that these values would be similar to past research. The purpose of this study was to reveal that the reason for the lack of convergence of the CTCM model is not due to the use of multiple dimensions, but rather the lack of manifest indicators. In fact, this study provides evidence for multiple dimensions, as opposed to one general performance factor.

An ancillary contribution of this study is demonstrating a way to subject ACs to the psychometric rigor associated with traditional measures. A common criticism of past AC research is the failure of AC researchers to pay careful consideration to the psychometric soundness of the dimensions assessed. In other words, AC researchers and practitioners appear to assume that the dimensions the AC was designed to measure are being appropriately measured. In order to investigate the representativeness of the behavioral checklists of the construct in which they are intended to measure, CFAs were performed. All of the parcels created from the behavioral checklist items significantly loaded on the intended dimensions. Although most checklists successfully represented the intended dimensions, there were three situations that were not very successful as deemed by commonly used fit criteria (Hu & Bentler, 1998, 1999). These three checklists were: leadership measured in role play, judgment measured in oral presentation, and organizational planning measured in oral presentation. Since all three of these examples have only four items on their checklists, we assume this lack of fit is based on the idiosyncrasies of the individual items. Specifically, the items did not have to be averaged and, therefore, are less reliable. To amend this, more checklist items could be added, which would then allow for averages to be made.

Implications

Examining the issues of the direct application of the CTCM model to AC PEDRs represents a key step forward in AC literature that has implications for future research and practice. First, AC research can now apply the use of multiple indicators to attain a convergent and admissible model that contains all exercises and all dimensions. Additionally, competing models of AC performance can be created in order to arrive at acceptable construct validity. More importantly, researchers now have empirical evidence that just as exercises may not be equivalent to methods (Lance, 2008), dimensions may not be interpreted as equivalents of traits (Hoffman & Meade, 2007). So it may seem that the assumptions of the CFA-MTMM are not met by ACs (Hoffman & Meade, 2007; Howard, 2008; Lance, 2008). Regardless of these discouraging results, AC research using different methodologies has provided positive validity evidence (Arthur et al., 2003; Meriac et al., 2008). This, combined with the results of the present study, provides evidence that the failure of the CTCM model is not representative of a bad AC, but an incorrect methodological application. Therefore, we can assume that both dimensions and exercises are relevant components of ACs.

Interestingly, assessee performance in dimensions is not necessarily consistent. In fact, when interpreting dimensional performance across exercises, consistencies and inconsistencies can be important and have been shown to be meaningful (Gibbons & Rupp, 2009). Therefore, it is not surprising that recent suggestions of dimension-free exercise-based ACs have been criticized (Arthur et al., 2008). Additionally, it seems dimensions are important components of ACs, not just conceptually, but theoretically (Arthur et al., 2003; Hoffman et al, in press; Meriac et al., 2008). In terms of feedback, dimension performance within exercises provides assessees more specific areas in which to focus their development.

Another important concept in examining the construct validity of ACs is the expectancy of dimensions to converge (i.e. demonstrate cross-situational consistency). If dimensions were intended to be consistent across exercises, current AC loadings are misleading. Most often, certain exercises measure certain dimensions better than others. Since some dimensions appear to be more easily observed in certain exercises, assuming dimensions are rated equivalently across exercises would seem counterintuitive. In fact, some believe that exercises provide different situations in which to elicit certain behaviors. The behaviors evoked are not necessarily identical across situations, since different situations place different demands on participants (Trait Activation Theory; see Lievens et al., 2006). This practice may prove beneficial since unique information about dimensions can be provided through separate situations.

Future Research

In using multiple manifest indicators, our results reveal a potential limitation of previous AC validity research. Specifically, previous AC research has assumed that the lack of convergence of the full CTCM model was indicative of the inappropriateness of the typical AC design. Instead, the failure to produce a proper solution when applying this model to AC data has in fact proved to be a methodological artifact not specific to ACs. Accordingly, our study contributes to the literature by providing evidence in favor of the construct validity of ACs. In light of support for the currently practiced AC, it may be the next step in AC research to frame a construct validity search to fit the already working AC tool.

Although there has been research against the retention of dimensions, it is important to acknowledge that dimensions contribute to the variance explained in ACs. Our results allow for a much improved interpretation of AC construct validity by providing an explanation for the lack of past success. Additionally, these results echo past dimension-supportive findings with comparative validity evidence (Hoffman, in press). With the importance of dimensions established, it may prove valuable to consider alternate conceptualizations of dimensions. For instance, future research should explore situationally specific dimensions (Hoffman & Meade, 2007), which assume that both differences in performance across exercises and differentiated performance within exercises exist. If performance dimensions are nested within exercises, performance in an AC may be attributable to both the dimensions and the situation in which those dimensions are evoked. This conceptualization may make developmental feedback more meaningful. For instance, instead of asking an individual to improve their oral presentation skills, feedback could be more directly given such that an individual is told organizational planning in writing is a strength, but organizational planning in presentations could be improved.

Another related area that future research should investigate is the possibility that dimensions may have never been intended to converge (Howard, 2008). According to Howard (2008), exercises are intended to measure different facets of a dimension; therefore, each exercise is not intended to capture the exact same information. If exercises are constructed to provide unique evidence about dimensions, it would be counterintuitive to expect these pieces of information to be consistent across different situations. Since both exercise factors and dimension factors have proved relevant in CFA analyses (Bycio, et al. 1987; Bowler & Woehr, 2006; Connelly et al., 2008), to get rid of either would be tantamount to ignoring a piece of the validity puzzle.

Researchers have determined that MTMM-based analyses may not be the most suitable approach to analyzing AC performance (Arthur et al., 2008; Howard, 2008). According to Lance (2008), "unquestionable adoption of the MTMM methodological platform to the AC PEDR arena may very well have misdirected researchers (a) to assume that exercise effects were undesirable sources of variance in PEDRs because they represent common method bias and (b) *away from* asking the right question, namely what constructs *are* being measured by AC PEDRs in the first place," (p. 90). Howard (2008) agrees with Lance (2008) but also adds that MTMM assumes exercises are designed to equally capture dimension information. Howard (2008) blames the application of MTMM on AC ratings for the misconception that dimensions should be consistent across exercises. "If each exercise is equally capable of measuring each dimension, why measure [a given dimension] five different times? Given the costs associated with

Assessment Centers, is this redundancy not an unnecessary waste of resources" (Howard, 2008, p. 99). She offered an alternative to the typical consistent dimension assumption by allowing certain dimensions to be considered "primary" in one exercise and "secondary" in another. This design supports the idea that exercises are designed to collect a variety of information about a dimension. As evidenced in this study, ACs may already be working in this fashion. In fact, it may be ideal for researchers and practitioners to consider an option that allows for unique information about dimensions to be collected from exercises.

If future researchers continue to utilize MTMM analyses, AC designers should consider assigning the same raters for each exercise. In fact, this is not an original thought. All raters should view all candidates in every exercise (Howard, 1997; Jones, 1992; Turnage & Muchinsky, 1982) in order to account for rater variance in addition to exercise and dimension variance. Although having different raters rate each exercise may be cheaper and faster than the same raters assessing performance in all exercises, it leads to a lack of explained variance (Howard, 1997; Jones, 1992). Since this particular AC used raters specific to each exercise, the exercise variance explained may be inflated due to the model's inability to take rater effects into account. Therefore, in order to prevent further inflation of exercise variance, raters should assess performance in all exercises.

Conclusion

The historically weak convergent validity of AC PEDRs does not appear to be the result of the use of worthless dimensions. Rather, with evidence of dimensions from this and other studies, the lack of convergent validity appears to be attributable to the sensitivity of the CTCM model under factor analyses. In fact, the convergence of the CTCM model can be achieved if the indicator-factor ratio is increased. As such, the past lack of success of AC construct validity is merely a consequence of a methodological artifact. Therefore, future research should seek other methodological alternatives. What does appear clear is that in order to fully understand assessee performance, AC researchers must utilize information from both exercises and dimensions.

REFERENCES

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003, Spring). 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., 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., Woehr, D. J., & Maldegan, 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.

Avison, W. R. (1978). Auxiliary theory and multitrait-multimethod validation: A review of two approaches. *Applied Measurement, 2,* 433-449. doi: 10.1177/014662167800200318

Bagozzi, R. P. (1993). Assessing construct validity in personality research: Applications to measures of self esteem. *Journal of Research in Personality, 27,* 49–87. doi: 10.1006/jrpe.1993.1005

Bagozzi, R. P., & Edwards, J. R. (1998, January). A general approach for representing constructs in organizational research. *Organizational Research Methods, 1,* 45-87.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246. doi: 10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

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. (2008). Back to basics of test construction and scoring. *Industrial and Organizational Psychology: Perspectives on Research and Practice, 1*, 131-133.

Brannick, M. T., & Spector, P. E. (1990). Estimation problems in the block-diagonal model of the multitrait-multimethod matrix. *Applied Psychological Measurement*, *14*, 325-339. doi: 10.1177/014662169001400401

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

Byrne, B. M., & Goffin, R. D. (1993). Modeling MTMM data from additive and multiplicative covariance structures: An audit of construct validity concordance. *Multivariate Behavioral Research*, *28,* 67-96. doi: 10.1207/s15327906mbr2801_5

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

Cattell, R. B. (1956). Validation and intensification of the Sixteen Personality Factor Questionnaire. *Journal of Clinical Psychology, 12,* 1956. pp. 205-214. doi: 10.1002/1097 4679(195607)12:3<205::AID-JCLP2270120302>3.0.CO;2-0

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

Connelly, B. S., Ones, D. S., Ramesh, A., & Goff, M.(2008). *Industrial and Organizational Psychology: Perspectives on Research 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.

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.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 3*, 493-511. doi: 10.1037/0021-9010.72.3.493

Graham, J. W. & Collins, N. L. (1991). Controlling correlational bias via confirmatory factor analysis of multitrait-multimethod data. *Multivariate Behavioral Research, 26,* 607-629. doi: 10.1207/s15327906mbr2604_3

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. doi: 10.1111/j.1744-6570.2002.tb00106.x

Hennessy, J., Mabey, B., & Warr, P. (1998) Assessment centre observation procedures: An experimental comparison of traditional, checklist, and coding methods. *International Journal of Selection and Assessment, 6,* 222-231. doi: 10.1111/1468-2389.00093

Hoffman, B.J. (in press). Research on Hybrid Approaches to Assessment Centers. In, Jackson, Lance, & Hoffman (Eds.) *Psychology of Assessment Centers.* Taylor Francis.

Hoffman, B. J., & Meade, A. W. (2007, April). *Invariance Tests as Assessment Center Construct Validity Evidence.* In B. J. Hoffman & C. E. Lance (chairs). *The Assessment Center Validity Paradox: Alternative Analytic and Design Methodologies*. Paper presented at the 22nd Annual Meeting of the Society for Industrial and Organizational Psychology, New York.

Hoffman, B. J., Melchers, K. G., Messal, C. B., Kleinmann, M., & Ladd, L. T. (in press). Exercises and dimensions are the currency of assessment centers. *Personnel Psychology.*

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

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

Jackson, D. J. 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

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

Jansen, P., & de Jongh, F., (1997). *Assessment Centres*. Chichester, England: John Wiley &

Sons.

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

Jöreskog K. G. (1974). Analyzing Psychological Data by Structural Analysis of Covariance matrices. In R. C. Atkinson, D. H. Krantz, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2, pp. 156). San Francisco: Freeman.

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

Kalleberg, Arne L.; Kluegel, James R. (1975, February). Analysis of the Multitrait-Multimethod Matrix: Some limitations and an alternative. *Journal* ***of*** *Applied Psychology, 60*, 1-9. doi: 10.1037/h0076267

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112,* 165-172. doi: 10.1037/0033 2909.112.1.165

Kishton, J.M, & Widaman, K.F., (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement, 54*, 757-765. doi: 10.1177/0013164494054003022

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., & Köller, 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.

Klimoski, R., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology, 30,* 243-260. doi: 10.1111/j.1744 6570.1987.tb00603.x

Kumar, A., & Dillon, W. R. (1992). An integrative look at the use of additive and multiplicative covariance structure models in the analysis of multitrait-multimethod data. *Journal of Marketing Research, 29*, 51-64. doi: 10.2307/3172492

Lance, C. (2008, March). Why assessment centers don't work the way they are suppose to. *Industrial and Organizational Psychology perspectives on science and practice, 1.*

Lance, C. E., Foster, M. R., Thoresen, J. D., & Gentry, W. A. (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. 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

Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychological Methods, 7*, 228-244. doi: 10.1037/1082-989X.7.2.228

Lievens, F. (2008). What does exercise-based assessment really mean? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 112-115.

Lievens, F. (2009). Assessment centres: A tale about dimensions, exercises, and dancing bears.

European Journal of Work and Organizational Psychology, 18, 102-121. doi: 10.1080/13594320802058997

Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006, March). 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. doi: 10.1037/0021-9010.91.2.247

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

Lowry, P. E., (1997). The assessment center process: New directions. *Journal of Social Behavior & Personality, 12*(5), 53-62.

MacCallum, R. C., Browne, M. W., Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait–multimethod data: Many problems and a few solutions *Applied Psychological Measurement, 13*, 335–361. doi: 10.1177/014662168901300402

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

Marsh, H. W., & Bailey, M. (1991, March). Confirmatory factor analysis of multitrait multimethod data: A comparison of alternative methods. *Applied Psychological Measurement, 15*, 47-70. doi: 10.1177/014662169101500106

Marsh, H.W., & Gouvernet, P.J. (1989). Multidimensional self-concepts and perceptions of control: Construct validation of responses by children. *Journal of Educational Psychology, 81,* 57-69. doi: 10.1037/0022-0663.81.1.57

Marsh, H. W., Hau, K., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in factor analysis. *Multivariate Behavioral Research, 33,* 181-220. doi: 10.1207/s15327906mbr3302

Meriac, J. P., Hoffman, B. J., Woehr, J., Fleisher, M. S. (2008, September). 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

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80,* 252–253 doi: 10.1037/h0035002

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

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 dimension. *Personnel Psychology, 43*, 71-84. doi: 10.1111/j.1744-6570.1990.tb02006.x

Sackett, P. R. (1987) Assessment centers and content validity: Some neglected issues. *Personnel Psychology, 40*, 11-25. coi: 10.1111/j.1744-6570.1987.tb02374.x

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

Sackett, P. R., & Tuzinski, K. A. (2001). The role of dimensions and exercises in assessment center judgments. In M. London (Ed.) *How people evaluate others in organizations.* (pp. 111-129). Mahwah, NJ: Erlbaum.

Sagie, A. & Magnezy, R. (1997) Assessor type number of distinguisable categories and assesstment cetnre construct validity. *Journal of Occupational and Organizational Psychology, 70,* 103-108.

Schmitt, N. (1978). Path analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 2,* 157-173. doi: 10.1177/014662167800200201

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, *10*, 1–22. doi: 10.1177/014662168601000101

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

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

Steiger, J. H. (1990). Structural model evaluation and modification. *Multivariate Behavioral Research*, *25,* 173-180. doi: 10.1207/s15327906mbr2502_4

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

Thornton, G. C., III. (1992). *Assessment centers and human resource management.* Reading, MA: Addison Wesley.

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

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

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10. doi: 10.1007/BF02291170

Turnage, J. J., & Muchinsky, P. M. (1982). Transsituational variability in human performance within assessment centers. *Organizational Behavior and Human Performance, 30*, 174 200. doi: 10.1016/0030-5073(82)90217-3

Velicer, W.F. Huckel, L.H., & Hansen C.E. (1989). A measurement model for measuring attitudes toward violence. *Personality and Social Psychology Bulletin, 15*, 349-364. doi: 10.1177/0146167289153006

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait multimethod data. *Applied Psychological Measurement*, 9, 1-26. doi: 10.1177/014662168500900101

Widaman, K. F. (1992). Multitrait-multimethod models in aging research. *Experimental Aging Research, 18,* 185-201.

Widaman, K.F., Gibbs, K.W., & Geary, D.C. (1987). Structure of adaptive behavior: I.

Replication across fourteen samples of nonprofoundly mentally retarded people. *American Journal of Mental Deficiency, 91*, 348-360.

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. doi: 10.1177/014920630302900206

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal. A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.

APPENDIX

Assessment Center Dimension Definitions

Dimension	Definition
Perception	The ability to identify key elements of a situation, the importance of these elements and their relationship to one another, including observation of relevant details and accurately recording information
Decisiveness	Involves the willingness/readiness to make timely decisions, take action, or commit one's self to a course of action and the willingness to accept responsibility for decisions
Judgment	Involves integrating a wide variety of information from written, oral and general sources, the development of alternative courses of action and making sound, logical decisions based on assumptions that reflect factual information; skill in this area is essential for both office activities and police field operations
Oral Communication	Involves skill in expressing one's self orally through the use of clear, well composed, and unambiguous statements. The quality of the speaking voice, use of facial and bodily gestures, and use of eye contact are included as part of the speaking component. The listening component of this dimension involves skill in picking out the most relevant aspects of what's being said by others, asking questions, giving feedback, and making demonstrable use of information gained by listening to others
Leadership	Involves getting ideas accepted and the direction, guidance and control of activities of other toward the accomplishment of tasks. This requires relating the needs of the department and the individual and monitoring the performance of the individual in order to provide assistance, to extend recognition, to discipline and/or to provide counseling. Additionally, this involves appropriate representation of view of employees and more senior managers to each other. Accomplishing this without generating resentment on anyone's part is an important aspect of this skill. The candidate must be able to get ideas accepted and influence others without alienating them. These results are achieved through delegation, control, and follow-up procedures
Organizational Planning	Involves establishing a course of action for self and/or others in order to accomplish a mission or work assignment. This involves planning the proper assignments of personnel and the appropriate allocation of resources and the organization of such personnel and resources
Written Communication	Involves skill in producing well organized, logical, and clearly written statements. Misuse of grammar, spelling, or punctuation are detrimental to the ratings only when these errors essentially change the intended meaning of the written statement or are excessive